FALL UPDATE
PDL PACKET

NEWSLETTER ON PDL ACTIVITIES AND EVENTS • FALL 2012
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
Riverbed Technology
Samsung Information Systems America
Seagate Technology
STEC, Inc.
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

JackRabbit: Improved Agility in Elastic Distributed Storage

Cipar, Xu, Krevat, Tumanov, Gupta, Kozuch & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-12-112, October 2012.

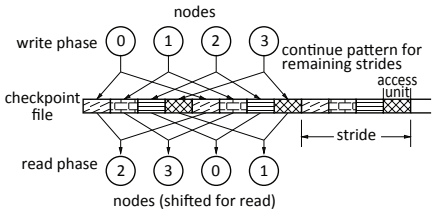
Elastic storage systems can be expanded or contracted to meet current demand, allowing servers to be turned off or used for other tasks. However, the usefulness of an elastic distributed storage system is limited by its agility: how quickly it can increase or decrease its number of servers. This paper describes an elastic storage system, called JackRabbit, that can quickly change its number of active servers. JackRabbit uses a combination of agility-aware offloading and reorganization techniques to minimize the work needed before deactivation or activation of servers. Analysis of real-world traces and experiments with the JackRabbit prototype confirm \sysname{}'s agility and show that it significantly reduces wasted server-time relative to state-of-the-art designs.

HPC Computation on Hadoop Storage with PLFS

Cranor, Polte & Gibson

Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-12-115, October 2012.

In this report we describe how we adapted the Parallel Log Structured Filesystem (PLFS) to enable HPC applications to be able read and write data from the HDFS cloud storage subsystem. Our enhanced version of PLFS provides HPC applications with the ability to concurrently write from multiple compute nodes into a single file stored in HDFS, thus allowing HPC applications to checkpoint. Our results show that HDFS combined with our PLFS HDFS I/O Store module is able to handle a concurrent write checkpoint workload generated by a benchmark with good performance.



Checkpoint benchmark operation.

Runtime Estimation of Stateless Exploration

Simsa, Bryant & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-12-113, October 2012.

In the past 15 years, stateless exploration, a collection of techniques for automated and systematic testing of concurrent programs, has experienced a wide-spread adoption. As stateless exploration moves into practice, becoming part of testing infrastructure of large-scale system developers, new pragmatic challenges are being identified. In this report we address the problem of accurate runtime estimation of stateless exploration by designing

continued on page 4

November 2012

Garth Gibson Keynote Speaker

Garth has been an invited keynote speaker at two conferences recently. His talk on "Storage Systems Issues for Shingled Magnetic Recording" opened the Storage System, Hard Disk and Solid State Technologies Summit, co-located with the Asia-Pacific Magnetic Recording Conference (APMRC), in Singapore, November 1, 2012. In September, he gave the SNIA SDC Keynote talk "Storage Systems for Shingled Disks" at the 2012 Storage Developer Conference in Santa Clara, CA.

October 2012

SCS Dissertation Award Winners Announced

Congratulations to the following PDL award winners, whose dissertations were chosen from among the many SCS produced last year. "Energy-efficient Data-intensive Computing with a Fast Array of Wimpy Nodes." by Vijay Vasudevan (Advisor: David Andersen) is one of two dissertations chosen for the top award. These dissertations will be nominated for the ACM Outstanding Dissertation Award. Receiving Honorable Mention is Duen Horng "Polo" Chau (Advisor: Christos Faloutsos) and his research on "Data Mining Meets HCI: Making Sense of Large Graphs."

October 2012

Carnegie Mellon University Repurposing Supercomputers From Los Alamos National Lab

The National Science Foundation, the New Mexico Consortium and Carnegie Mellon University joined forces to launch PRObE, a one-of-a-kind supercomputer research center using a cluster of 2,048 recently retired computers. The Tribune-Review reports, "Although the main facility will remain in Los Alamos, Carnegie Mellon's Parallel Data Lab in Pittsburgh will house two smaller centers." Garth Gibson, who collaborated on the project, described the Pittsburgh facility as a 'staging cluster,' and will allow researchers to perform small experiments and demonstrate to the PRObE committee that they're ready to request time on the facility in Los Alamos, known as Kodiak."

--info from the Pittsburgh Tribune-Review and First Bell Engineering and Technology News, Oct. 23, 2012

October 2012

Christos Faloutsos and Team win Big Data Award

The National Science Foundation (NSF), with support from the National Institutes of Health (NIH), recently announced nearly $15 million in new Big Data fundamental research projects. These awards aim to develop new tools and methods to extract and use knowledge from collections of large data sets to accelerate progress in science and engineering research and innovation.

The eight winning projects announced run the gamut of scientific techniques for big data management, new data analytic approaches, and e-science collaboration environments with possible future applications in a variety of fields, such as physics, economics and medicine.

Christos Faloutsos (PI), Tom Mitchell (co-PI) and their team proposed the successful project "BIGDATA: Mid-Scale: DA: Collaborative Research: Big Tensor Mining: Theory, Scalable Algorithms and Applications." The objective of the project is to develop theory and algorithms to tackle the complexity of language processing, and to develop methods that approximate how the human brain works in processing language. The research also promises better algorithms for search engines, new approaches to understanding brain activity, and better recommendation systems for retailers.

--from NSF Press Release 12-187

September 2012

Congratulations Henggang and Xiaowen!

Best wishes to Henggang Cui and Xiaowen Ding, who were married on September 28 in a local civil ceremony. They took pictures in China to celebrate before coming to Pittsburgh, and plan on having a full wedding at a later date.

July 2012

PDL Alum Receives ACM SIGKDD Dissertation Honors

Lei Li, now a post-doctoral researcher at the University of California, Berkeley, who earned his PhD in computer science in 2011, was the runner up for the prestigious 2012 Doctoral Dissertation Award from the Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (ACM SIGKDD).

In his dissertation, "Fast Algorithms for Mining Co-evolving Time Series," Li developed novel algorithms for forecasting, clustering and missing-value imputation for time sequences in a broad spectrum of settings, from

continued on page 3

continued from page 2

motion-capture sequences to data-center monitoring. Christos Faloutsos, professor of computer science, was Li's advisor.

The ACM SIGKDD dissertation award is the highest distinction for a PhD in the field.

--from CMU School of Computer Science News, July 27, 2012

May 2012

Faloutsos to Receive Honorary Degree from Aristotle University

The Aristotle University of Thessaloniki, the largest university in Greece, will award an honorary doctorate degree to Christos Faloutsos, professor of computer science. The title of Doctor Honoris Causa will be conferred to Faloutsos during a May 30 convocation. During convocations, Faloutsos will present a convocation address on "Mining Large Social Networks: Patterns and Anomalies."

Faloutsos' research interests include data mining for graphs and streams, fractals, database performance and indexing for multimedia and bio-informatics data, and his cross-disciplinary work is widely and regularly cited.

--from CMU News May 30, 2012

May 2012

PDL Alum Ryan Johnson Wins SIGMOD Jim Gray Doctoral Dissertation Award

Congratulations to Ryan Johnson, who has received the very prestigious SIGMOD Jim Gray Doctoral Dissertation Award! SIGMOD has established the annual SIGMOD Jim Gray Doctoral Dissertation Award to recognize excellent research by doctoral candidates in the database field. This award, which was previously known as the SIGMOD Doctoral Dissertation Award, was renamed in 2008 with the unanimous approval of ACM Council in honor of Dr. Jim Gray.

Ryan received the award for his PhD thesis titled "Scalable Storage Managers for the Multicore Era".

May 2012

Welcome Oscar!

Congratulations to Jim and Melanie Cipar, who welcomed their first child, Oscar David, on May 20th. He was 9lbs and 22 inches, with reddish-blonde hair and blue eyes.

May 2012

Priya's Student wins Alumni Award for Undergraduate Excellence in CS

We are pleased to announce that this year's recipient of the Alumni Award for Undergraduate Excellence in Computer Science is Nikhil Khadke, for his work entitled "Transparent System Call Based Performance Debugging for Cloud Computing." Nikhil is advised by Priya Narasimhan.

May 2012

David Andersen Collaborates with Intel ISTC-CC Researchers to Win JouleSort Competition!

A team from the ISTC for Cloud Computing—Babu Pillai, Michael Kaminsky, Mike Kozuch (Intel Labs), and Dave Andersen (CMU)—were announced winners in 3 categories of the 2012 JouleSort competition, setting new records for fewest joules needed to sort 108, 109, and 1010 records. The team used an Intel Core i7-2700K desktop processor, coupled with 16 Intel 710 Series SSDs to beat existing energy efficiency records in the 10GB, 100GB, and 1TB categories by 2.6% (their record from last year), 33%, and 729%, respectively.

April 2012

Wolf Richter Receives Alan J. Perlis SCS Student Teaching Award

Congratulations to Wolfgang Richter, who has received the Alan J. Perlis Graduate Student Teaching Award for 2012. The awards, for both graduate and undergraduate teaching assistants, are based on student nominations, recommendation letters and reviews, and honors the students who have shown the highest degree of excellence and dedication as teaching assistants.

June 2011

Tumanov Earns Canada Graduate Scholarship

ECE Ph.D. student Alexey Tumanov has earned a National Sciences and Engineering Research Council of Canada (NSERC) Alexander Graham Bell Canada Graduate Scholarship (CGS-D) that will support his research on cloud computing with Jatras Professor of ECE Greg Ganger. The prestigious award is presented to Canadian citizens or permanent residents pursuing doctor's degrees, and is based on the applicant's academic excellence, research ability and potential, and communications, interpersonal and leadership abilities. The NSERC is the Canadian equivalent of the NSF, and only the top tier of post-graduate scholarship recipients earn the CGS-D.

--from ECE News Online, June 17, 2011

continued from page 1

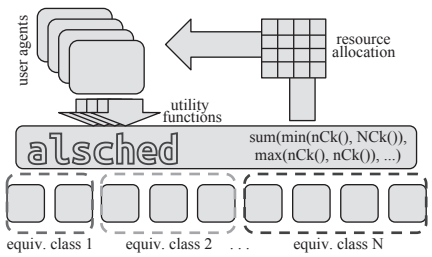
techniques that address the non-linear nature through which modern stateless exploration techniques enumerate the state space of possible test executions and evaluate our techniques on a collection of exploration traces from a real-world deployment at Google.

alsched: Algebraic Scheduling of Mixed Workloads in Heterogeneous Clouds

Tumanov, Cipar, Kozuch & Ganger

3rd ACM Symposium on Cloud Computing. Oct. 14-17, 2012 - San Jose, CA.

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



alsched System Model

Landslide: Systematic Exploration for Kernel-Space Race Detection

Blum & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-12-114, October 2012.

Systematic exploration is an approach to finding race conditions by deterministically executing many thread interleavings and identifying which ones expose bugs. Current techniques are suitable for testing user-space programs, but are inadequate for testing operating system kernels. Testing kernel-level code necessitates understanding the kernel's design in order to effectively control nondeterminism and achieve reasonable state-space reduction. We present Landslide, a systematic exploration tool for finding races in kernels. Landslide makes use of user-provided configuration to enable efficient exploration and testing of meaningful interleavings. The user instruments the kernel to inform Landslide of important concurrency events, and configures Landslide's search to de-emphasize irrelevant kernel components. This combines the user's design knowledge with Landslide's ability to explore large state spaces. Our experience with Landslide shows that a tool built for this usage pattern is capable of identifying otherwise-overlooked kernel-space races.

A Case for Scaling HPC Metadata Performance through De-specialization

Patil, Ren & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-12-111, October 2012.

Modern cluster file systems provide highly scalable I/O bandwidth along the data path by enabling highly parallel access to file data. Unfortunately metadata scaling is lagging behind data scaling. We propose a file system design that inherits the scalable data bandwidth of existing cluster file systems and adds support for distributed and high-performance metadata operations. Our key idea is to integrate a distributed indexing mechanism with general-purpose optimized on-disk metadata store. To demonstrate the feasibility of our approach, we implemented a prototype middleware layer using the FUSE file system and evaluated it on 64-node cluster. Preliminary results show promising scalability and performance: the single- node local metadata store was 10X faster than modern local file systems and the distributed middleware metadata service scaled well with a peak performance of 190,000 file creates per second on a 64-server configuration.

HAT: Heterogeneous Adaptive Throttling for On-Chip Networks

Chang, Ausavarungnirun, Fallin & Mutlu

SBAC-PAD 2012, New York, NY, October 24-26, 2012.

The network-on-chip (NoC) is a primary shared resource in a chip multiprocessor (CMP) system. As core counts continue to increase and applications become increasingly data-intensive, the network load will also increase, leading to more congestion in the network. This network congestion can degrade system performance if the network load is not appropriately controlled. Prior works have proposed sourcethrottling congestion control, which limits the rate at which new network traffic (packets) enters the NoC in order to reduce congestion and improve performance. These prior congestion control mechanisms have shortcomings that significantly limit their performance: either 1) they are not application-aware, but rather throttle all applications equally regardless of applications' sensitivity to latency, or 2) they are not network-load-aware, throttling according to application characteristics but sometimes

continued on page 5

continued from page 4

under- or over-throttling the cores.

In this work, we propose Heterogeneous Adaptive Throttling, or HAT, a new source-throttling congestion control mechanism based on two key principles: application-aware throttling and network-load-aware throttling rate adjustment. First, we observe that only network-bandwidth-intensive applications (those which use the network most heavily) should be throttled, allowing the other latency-sensitive applications to make faster progress without as much interference. Second, we observe that the throttling rate which yields the best performance varies between workloads; a single, static, throttling rate under-throttles some workloads while over-throttling others. Hence, the throttling mechanism should observe network load dynamically and adjust its throttling rate accordingly. While some past works have also used a closed-loop control approach, none have been application-aware. HAT is the first mechanism to combine application-awareness and network-load-aware throttling rate adjustment to address congestion in a NoC.

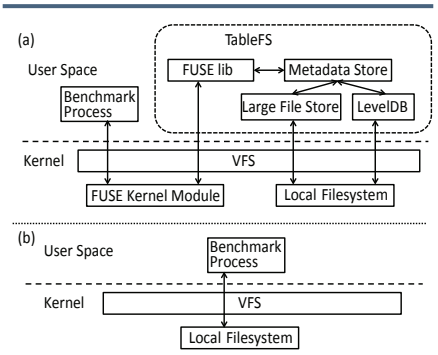
We evaluate HAT using a wide variety of multiprogrammed workloads on several NoC-based CMP systems with 16-, 64-, and 144-cores and compare its performance to two state-of-the-art congestion control mechanisms. Our evaluations show that HAT consistently provides higher system performance and fairness than prior congestion control mechanisms.

TABLEFS: Embedding a NoSQL Database Inside the Local File System

Ren & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-12-110, September 2012.

File systems that manage magnetic disks have long recognized the importance of sequential allocation and large



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

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

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

Row Buffer Locality Aware Caching Policies for Hybrid Memories

Yoon, Meza, Ausavarungnirun, Harding & Mutlu

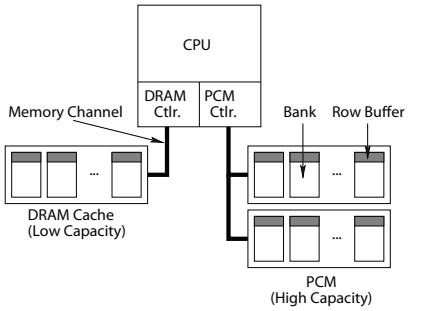
Proceedings of the 30th IEEE International Conference on Computer Design (ICCD 2012), Montreal, Quebec, Canada, September 2012. Best paper award in Computer Systems and Applications track.

Phase change memory (PCM) is a promising technology that can offer higher capacity than DRAM. Unfortunately, PCM's access latency and energy are higher than DRAM's and its endurance is lower. Many DRAM-PCM hybrid memory systems use DRAM as a cache to PCM, to achieve the low access latency and energy, and high endurance of DRAM, while taking advantage of PCM's large capacity. A key question is what data to cache in DRAM to best exploit the advantages of each technology while avoiding its disadvantages as much as possible.

We propose a new caching policy that improves hybrid memory performance and energy efficiency. Our observation is that both DRAM and PCM banks employ row buffers that act as a cache for the most recently accessed memory row. Accesses that are row buffer hits incur similar latencies (and energy consumption) in DRAM and PCM, whereas accesses that are row buffer misses incur longer latencies (and higher energy consumption) in PCM. To exploit this, we devise a policy that avoids accessing in PCM data that frequently causes row buffer misses because such accesses are costly in terms of both latency and energy. Our policy tracks the row buffer miss

continued on page 6

continued from page 5



Hybrid memory system organization.

counts of recently used rows in PCM, and caches in DRAM the rows that are predicted to incur frequent row buffer misses. Our proposed caching policy also takes into account the high write latencies of PCM, in addition to row buffer locality.

Compared to a conventional DRAM-PCM hybrid memory system, our row buffer locality-aware caching policy improves system performance by 14% and energy efficiency by 10% on data-intensive server and cloud-type workloads. The proposed policy achieves 31% performance gain over an all-PCM memory system, and comes within 29% of the performance of an all- DRAM memory system (not taking PCM's capacity benefit into account) on evaluated workloads.

Heterogeneity and Dynamicity of Clouds at Scale: Google Trace Analysis

Reiss, Tumanov, Ganger, Katz & Kozuch

3rd ACM Symposium on Cloud Computing. October 14th-17th, 2012 - San Jose, CA.

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

Indexing and Fast Near-Matching of Billions of Astronomical Objects

Fu, Fink, Gibson & Carbonell

Proceedings of the Fourth Workshop on Interfaces and Architecture for Scientific Data Storage, 2012 (IASDS12). September 24, 2012, Beijing, China.

When astronomers analyze sky images, they need to identify the newly observed celestial objects in the catalog of known objects. We have developed a technique for indexing catalogs, which supports fast retrieval of closely matching catalog objects for every object in new images. It allows processing of a sky image in less than a second, and it scales to catalogs with billions of objects.

A Proof of Correctness for Egalitarian Paxos

Moraru, Andersen & Kaminsky

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-12-109. September 2012.

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.

How Does Your Password Measure Up? The Effect of Strength Meters on Password Creation

Ur, Kelley, Komanduri, Lee, Maass, Mazurek, Passaro, Shay, Vidas, Bauer, Christin & L. Cranor

In the 2012 USENIX Security Symposium, August 2012.

To help users create stronger text-based passwords, many web sites have deployed password meters that provide visual feedback on password strength. Although these meters are in wide use, their effects on the security and usability of passwords have not been well studied. We present a 2,931-subject study of password creation in the presence of 14 password meters. We found that meters with a variety of visual appearances led users to create longer passwords. However, significant increases in resistance to a password-cracking algorithm were only achieved using meters that scored passwords stringently. These stringent meters also led participants to include more digits, symbols, and uppercase letters. Password meters also affected the act of password creation. Participants who saw stringent meters spent longer creating their password and were more likely to change their password while entering it, yet they were

continued on page 7

continued from page 6

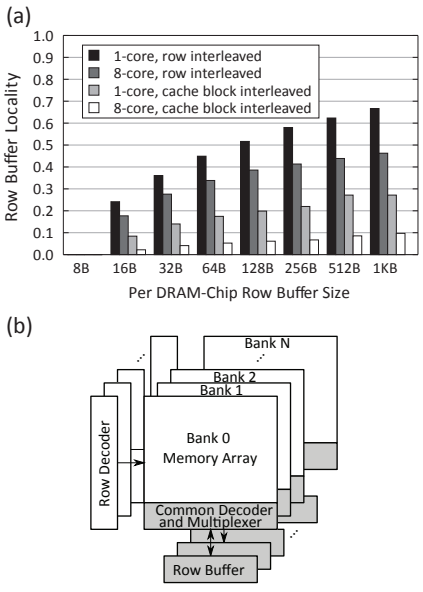
also more likely to find the password meter annoying. However, the most stringent meter and those without visual bars caused participants to place less importance on satisfying the meter. Participants who saw more lenient meters tried to fill the meter and were averse to choosing passwords a meter deemed "bad" or "poor." Our findings can serve as guidelines for administrators seeking to nudge users towards stronger passwords.

A Case for Small Row Buffers in Non-Volatile Main Memories

Meza, Li & Mutlu

Proceedings of the 30th IEEE International Conference on Computer Design (ICCD 2012), Poster Session, Montreal, Quebec, Canada, September 2012.

DRAM-based main memories have read operations that destroy the read data, and as a result, must buffer large amounts of data on each array access to keep chip costs low. Unfortunately, system-level trends such as increased memory contention in multi-core architectures and data mapping schemes that improve memory parallelism lead to only a small amount of the buffered data to be accessed. This makes buffering large amounts of data on every memory array access energy-inefficient; yet organizing DRAM chips to buffer small amounts of data is costly, as others have shown.



Row size affects row locality (a); our NVM architecture (b).

Emerging non-volatile memories (NVMs) such as PCM, STT-RAM, and RRAM, however, do not have destructive read operations, opening up opportunities for employing small row buffers without incurring additional area penalty and/or design complexity. In this work, we discuss and evaluate architectural changes to enable small row buffers at a low cost in NVMs. We find that on a multi-core system, reducing the row buffer size can greatly reduce main memory dynamic energy compared to a DRAM baseline with large row sizes, without greatly affecting endurance, and for some NVM technologies, leads to improved performance.

Egalitarian Paxos

Moraru, Andersen & Kaminsky

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-12-108. July 2012.

This paper describes the design and implementation of Egalitarian Paxos (EPaxos), a new distributed consensus algorithm based on Paxos. EPaxos achieves two 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; and (2) uniform load balancing across all replicas—no replicas experience higher load because they have special roles. Egalitarian Paxos is to our knowledge the first distributed consensus protocol to achieve both 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.

Staged Memory Scheduling: Achieving High Performance and Scalability in Heterogeneous Systems

Ausavarungnirun, Chang, Subramanian, Loh & Mutlu

The 39th International Symposium on Computer Architecture (ISCA), Portland, Oregon, June 9-13th, 2012.

When multiple processor (CPU) cores and a GPU integrated together on the same chip share the off-chip main memory, requests from the GPU can heavily interfere with requests from the CPU cores, leading to low system performance and starvation of CPU cores. Unfortunately, state-of-the-art application-aware memory scheduling algorithms are ineffective at solving this problem at low complexity due to the large amount of GPU traffic. A large and costly request buffer is needed to provide these algorithms with enough visibility across the global request stream, requiring relatively complex hardware implementations.

This paper proposes a fundamentally new approach that decouples the memory controller's three primary tasks into three significantly simpler structures that together improve system performance and fairness, especially in integrated CPU-GPU systems. Our three-stage memory controller first groups requests based on row-buffer locality. This grouping allows the second stage to focus only on inter-application request scheduling. These two stages enforce high-level policies regarding performance and fairness, and therefore the last stage consists of simple per-bank FIFO queues (no further command reordering within each bank) and straightforward logic

continued on page 10

DISSERTATION ABSTRACT: Chaining for Flexible and High-Performance Key-Value Systems

Amar Phanishayee

Carnegie Mellon University SCS Ph.D. Dissertation, September 17, 2012

Distributed key-value (KV) systems are a critical part of the infrastructure at many large sites such as Amazon, Facebook, Google, and Twitter. Unfortunately, the ecosystem of these KV systems is a mess—no one existing system meets the needs of all applications. Systems designers worry about running multiple stores from different codebases, vendors, and so on, each optimized for certain application requirements and hardware configuration. We argue that having systems designers worry about running multiple stores from different codebases, vendors, and so on, each optimized for certain application requirements and hardware configuration, is unreasonable and unnecessary.

This dissertation proposes a key-value architecture using a generalization of chain-based replication which can be easily configured to support many points along the KV design continuum. First, we present a new replication protocol, Ouroboros, which extends chain-based replication to allow node additions to any part of the replica chain, minimize blocking during node additions and deletions, and guarantee provably strong data consistency. We use Ouroboros in the implementation of a distributed key-value storage system, FAWN-KV, designed with the goal of supporting the three key properties of fault tolerance, high performance, and generality. Second, we present a generalization of chain-based replication to effectively support a wide range of application requirements using four simple knobs: (a) replica type; (b) replication factor; (c) update mechanism between replicas; and (d) query node selection. We describe Flex-KV, that extends Ouroboros with this generalization. Flex-KV can support DRAM, Flash, and disk-based storage; can act as an unreliable cache or a durable store; and can offer strong or weak data consistency. The value of such a system goes beyond ease-of-use: While exploring these dimensions of durability, consistency, and availability, we find new choices for system designs, such as a cache-consistent memcached, that offer some applications a better balance of performance and cost than was previously available.

Greg Ganger welcomes our industry visitors to the 2012 PDL Spring Visit Day.

DISSERTATION ABSTRACT: Mining and Modeling Real-world Networks: Patterns, Anomalies, and Tools

Leman Akoglu

Carnegie Mellon University SCS Ph.D. Dissertation, August 22, 2012

Large real-world graph (a.k.a. network, relational) data are omnipresent, in online media, businesses, science, and the government. Analysis of these massive graphs is crucial, in order to extract descriptive and predictive knowledge with many commercial, medical, and environmental applications. In addition to its general structure, knowing what stands out, i.e. anomalous or novel, in the data is often at least, or even more important and interesting. In this thesis, we build novel algorithms and tools for mining and modeling large-scale graphs, with a focus on:

Brian Hirano of Oracle speaks on "Advocating Efficient Scheduling Support in Hardware for Software" at a special PDL Consortium Speaker Series.

(1) Graph pattern mining: we discover surprising patterns that hold across diverse real-world graphs, such as the "fortification effect" (e.g. the more donors a candidate has, the super-linearly more money s/he will raise), dynamics of connected components over time, and power-laws in human communications,

(2) Graph modeling: we build generative mathematical models, such as the RTG model based on "random typing" that successfully mimics a long list of properties that real graphs exhibit,

(3) Graph anomaly detection: we develop a suite of algorithms to spot abnormalities in various conditions; for (a) plain weighted graphs, (b) binary and categorical attributed graphs, (c) time-evolving graphs, and (d) sense-making and visualization of anomalies.

DISSERTATION ABSTRACT: Understanding and Managing Propagation on Large Networks—Theory, Algorithms, and Models

B. Aditya Prakash

Carnegie Mellon University SCS Ph.D. Dissertation, September 18, 2012

How do contagions spread in population networks? What happens if the networks are dynamic? Which hospitals should we give vaccines to, for maximum effect? How to detect sources of

continued on page 9

continued from page 8

rumors on Twitter/Facebook?

These questions and many others such as which group should we market to, for maximizing product penetration, how quickly news travels in online media and how the relative frequencies of competing tasks evolve are all related to propagation/cascade-like phenomena on networks.

In this thesis, we present novel theory, algorithms and models for propagation processes on large static and dynamic networks, focusing on:

1. Theory: We tackle several fundamental questions like determining if there will be an epidemic, given the underlying networks and virus propagation models and predicting who-wins when viruses (or memes or products etc.) compete. We give a unifying answer for the threshold based on eigenvalues, and prove the surprising "winner-takes-all" result and other subtle phase-transitions for competition among viruses.

2. Algorithms: Based on our analysis, we give dramatically better algorithms for important tasks like effective immunization and reliably detecting culprits of epidemics. Thanks to our carefully designed algorithms, we achieve 6x fewer infections on real hospital patient-transfer graphs while also being significantly faster than other competitors (up to 30,000x).

3. Models: Finally using our insights, we study numerous datasets to develop powerful general models for information diffusion and competing species in a variety of situations. Our models unify earlier patterns and results, yet being succinct and enable challenging tasks like trend forecasting, spotting outliers and answering 'what-if' questions.

Our inter-disciplinary approach has led to many discoveries in this thesis, with broad applications spanning areas like public health, social media, product marketing and networking. We are arguably the first to present a systematic study of propagation and immunization of single as well as multiple viruses on arbitrary, real and time-varying networks as the vast majority of the literature focuses on structured topologies, cliques, and related un-realistic models.

Raja Sambasivan describes his research on "Diagnosing Performance Changes by Comparing Request Flows" to Athicha Muthitacharoen, Google and Brian Mueller, Panasas at a PDL Spring Visit Day poster session.

THESIS PROPOSAL:
Trading Latency for Freshness in Storage Systems

Jim Cipar, SCS

October 2011

Many storage systems have to provide extremely high throughput updates and low latency read queries. In practice, system designs that provide those capabilities often face a trade-off between query latency, efficiency, and result freshness. In my dissertation, I will argue that systems should be designed to allow for a per-query configuration of this trade-off. I will use two case studies to demonstrate the value of doing so. The first is LazyBase, a database designed for high-throughput ingest of observational data. The second is LazyTables, a shared data structure designed to support parallel machine learning applications. In both cases, the term "Lazy" refers to the systems' procrastination: waiting to apply updates until they can be executed as efficiently as possible. This design decision creates the potential for staleness in the data, hence the need for studying the trade-off between freshness and performance. Additionally, I will describe a number of other applications where this trade-off is potentially useful in system design.

M.S. THESIS:
Landslide: Systematic Dynamic Race Detection in Kernel-space

Ben Blum, SCS

May 10, 2012

Systematic exploration is an approach to finding race conditions by deterministically executing every possible interleaving of thread transitions and identifying which ones expose bugs. Current systematic exploration techniques are suitable for testing user-space programs, but are inadequate for testing kernels, where the testing framework's control over concurrency is more complicated. We present Landslide, a systematic exploration tool for finding races in kernels. Landslide targets Pebbles, the kernel specification that students implement in the undergraduate Operating Systems course at Carnegie Mellon University (15- 410). We discuss the techniques Landslide uses to address the general challenges of kernel-level concurrency, and we evaluate its effectiveness and usability as a debugging aid. We show that our techniques make systematic testing in kernel-space feasible, and that Landslide is a useful tool for doing so in the context of 15-410.

Bill Courtright and Jim Zelenka (PDL Alum, Panasas) enjoy the sun while they visit about storage at the 2012 PDL Spring Visit Day.

continued from page 7

that deals only with low-level DRAM commands and timing.

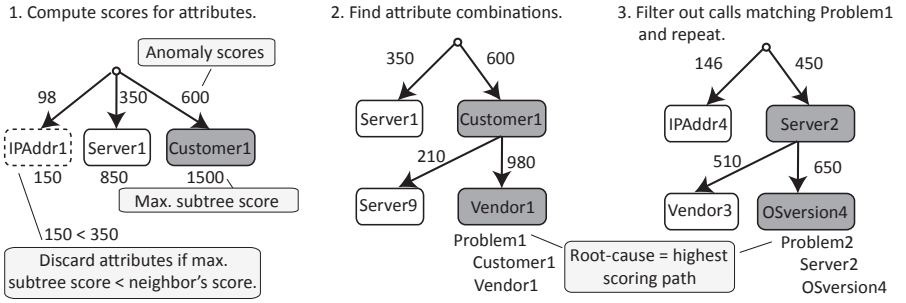
We evaluate the design trade-offs involved in our Staged Memory Scheduler (SMS) and compare it against three state-of-the-art memory controller designs. Our evaluations show that SMS improves CPU performance without degrading GPU frame rate beyond a generally acceptable level, while being significantly less complex to implement than previous application-aware schedulers. Furthermore, SMS can be configured by the system software to prioritize the CPU or the GPU at varying levels to address different performance needs.

Draco: Statistical Diagnosis of Chronic Problems in Large Distributed Systems

Kavulya, Daniels, Joshi, Hiltunen, Gandhi & Narasimhan.

IEEE/IFIP Conference on Dependable Systems and Networks (DSN), June 2012.

Chronics are recurrent problems that often fly under the radar of operations teams because they do not affect enough users or service invocations to set off alarm thresholds. In contrast with major outages that are rare, often have a single cause, and as a result are relatively easy to detect and diagnose quickly, chronic problems are elusive because they are often triggered by complex conditions, persist in a system for days or weeks, and coexist with other problems active at the same time. In this paper, we present Draco, a scalable engine to diagnose chronics that addresses these issues by using a "topdown" approach that starts by heuristically identifying user interactions that are likely to have failed, e.g., dropped calls, and drills down to identify groups of properties that best explain the difference between failed and successful interactions by using a scalable Bayesian learner. We have deployed Draco in production for the VoIP operations of a major ISP. In addition to providing examples of chronics that Draco has helped identify, we show via a comprehensive evaluation on production data that Draco provided 97% coverage, had fewer than 4% false positives, and outperformed state-of-the-art diagnostic techniques by up to 56% for complex chronics.



Draco uses an iterative Bayesian approach to rank combinations of attributes most correlated with the problem.

Light-weight Black-box Failure Detection for Distributed Systems

Tan, Kavulya, Gandhi & Narasimhan

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-12-107. July 2012.

Diagnosing failures in distributed systems is challenging, as modern datacenters run a variety of applications and systems. Current techniques for detecting failures often require training, have limited scalability, or are not intuitive to sysadmins. We present LFD, a lightweight and scalable technique for diagnosing performance problems in distributed systems using only correlations of operating system metrics collected transparently. The LFD fault detection algorithm is based on our hypothesis of server application behavior, and hence does not require training, and can perform failure detection with complexity linear in the number of nodes, with results that are intuitively interpretable by sysadmins. Further, with some training, LFD-DT uses decision-trees to diagnose the category of a problem that has previously been seen. We further show that LFD is versatile, and can diagnose faults in Hadoop MapReduce systems and on multi-tier web request systems, and show how LFD is intuitive to sysadmins.

Correct Horse Battery Staple: Exploring the Usability of System-assigned Passphrases

Shay, Kelley, Komanduri, Mazurek, Ur, Vidas, Bauer, Christin & L. Cranor

In SOUPS 2012: Symposium on Usable Privacy and Security, July 2012.

Users tend to create passwords that are easy to guess, while system-assigned passwords tend to be hard to remember. Passphrases, space-delimited sets of natural language words, have been suggested as both secure and usable for decades. In a 1,476-participant online study, we explored the usability of 3- and 4-word system-assigned passphrases in comparison to system-assigned passwords composed of 5 to 6 random characters, and 8-character system-assigned pronounceable passwords. Contrary to expectations, system- assigned passphrases performed similarly to system-assigned passwords of similar entropy across the usability metrics we examined. Passphrases and passwords were forgotten at similar rates, led to similar levels of user difficulty and annoyance, and were both written down by a majority of participants. However, passphrases

continued on page 11

continued from page 10

took significantly longer for participants to enter, and appear to require error-correction to counteract entry mistakes. Passphrase usability did not seem to increase when we shrunk the dictionary from which words were chosen, reduced the number of words in a passphrase, or allowed users to change the order of words.

Exact and Approximate Computation of a Histogram of Pairwise Distances between Astronomical Objects

Fu, Fink, Gibson & Carbonell

First Workshop on High Performance Computing in Astronomy (AstroHPC 2012), held in conjunction with the 21st International ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC 2012), June 18-19, 2012, Delft, the Netherlands.

We compare several alternative approaches to computing correlation functions, which is a cosmological application for analyzing the distribution of matter in the universe. This computation involves counting the pairs of galaxies within a given distance from each other and building a histogram that shows the dependency of the number of pairs on the distance.

The straightforward algorithm for counting the exact number of pairs has the $O(n^2)$ time complexity, which is unacceptably slow for most astronomical and cosmological datasets, which include billions of objects. We analyze the performance of several alternative algorithms, including the exact computation with an $O(n^{5/3})$ average running time, an approximate computation with linear running time, and another approximate algorithm with sub-linear running time, based on sampling the given dataset and computing the correlation functions for the samples. We compare the accuracy of the described algorithms and analyze the tradeoff between their accuracy and running time. We also propose a novel hybrid approximation algorithm, which outperforms each other technique.

Automated Diagnosis without Predictability is a Recipe for Failure

Sambasivan & Ganger

Proceedings of the 4th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud '12), June 12-13, 2012, Boston, MA. Supersedes Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-11-101.

Automated management is critical to the success of cloud computing, given its scale and complexity. But, most systems do not satisfy one of the key properties required for automation: predictability, which in turn relies upon low variance. Most automation tools are not effective when variance is consistently high. Using automated performance diagnosis as a concrete example, this position paper argues that for automation to become a reality, system builders must treat variance as an important metric and make conscious decisions about where to reduce it. To help with this task, we describe a framework for reasoning about sources of variance in distributed systems and describe an example tool for helping identify them.

MinBD: Minimally-Buffered Deflection Routing for Energy-Efficient Interconnect

Fallin, Nazario, Yu, Chang, Ausavarungnirun & Mutlu

In NOCS 2012, Lyngby, Denmark, May 2012. (Best Paper Award Nominee)

A conventional Network-on-Chip (NoC) router uses input buffers to store in-flight packets. These buffers improve performance, but consume significant power. It is possible to bypass these buffers when they are empty, reducing dynamic power, but static buffer power, and dynamic power when buffers are utilized, remain. To improve energy efficiency, bufferless deflection routing removes input buffers, and instead uses deflection (misrouting) to resolve contention. However, at high network load, deflections cause unnecessary network hops, wasting power and reducing performance.

In this work, we propose a new NoC router design called the minimally-buffered deflection (MinBD) router. This router combines deflection routing with a small "side buffer," which is much smaller than conventional input buffers. A MinBD router places some network traffic that would have otherwise been deflected in this side buffer, reducing deflections significantly. The router buffers only a fraction of traffic, thus making more efficient use of buffer space than a router that holds every flit in its input buffers. We evaluate MinBD against input-buffered routers of various sizes that implement buffer bypassing, a bufferless router, and a hybrid design, and show that MinBD is more energy-efficient than all prior designs, and has performance that approaches the conventional input-buffered router with area and power close to the bufferless router.

Hadoop's Adolescence: A Comparative Workload Analysis from Three Research Clusters

Ren, Kwon, Balazinska, Howe

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-12-106. June 2012.

We analyze Hadoop workloads from three different research clusters from an application-level perspective, with two goals: (1) explore new issues in application patterns and user behavior and (2) understand key performance challenges related to IO and load balance. Our analysis suggests that Hadoop usage is still in its adoles-

continued on page 12

continued from page 11

cence. We see underuse of Hadoop features, extensions, and tools as well as significant opportunities for optimization. We see significant diversity in application styles, including some "interactive" workloads, motivating new tools in the ecosystem. We find that some conventional approaches to improving performance are not especially effective and suggest some alternatives. Overall, we find significant opportunity for simplifying the use and optimization of Hadoop, and make recommendations for future research.

SkyeFS: Distributed Directories using Giga+ and PVFS

Chivetta, Patil & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-12-104, May 2012.

There is growing set of large-scale data-intensive applications that require file system directories to store millions to billions of files in each directory and to sustain hundreds of thousands of concurrent directory operations per second. Unfortunately, most cluster file systems are unable to provide this level of scale and parallelism. In this research, we show how the GIGA+ distributed directory algorithm, developed at CMU, can be applied to a real-world cluster file system. We designed and implemented a user-level file system, called SkyeFS, that efficiently layers GIGA+ on top of the PVFS cluster file system. Our experimental evaluation demonstrates how an optimized interposition layer can help PVFS achieve the desired scalability for massive file system directories.

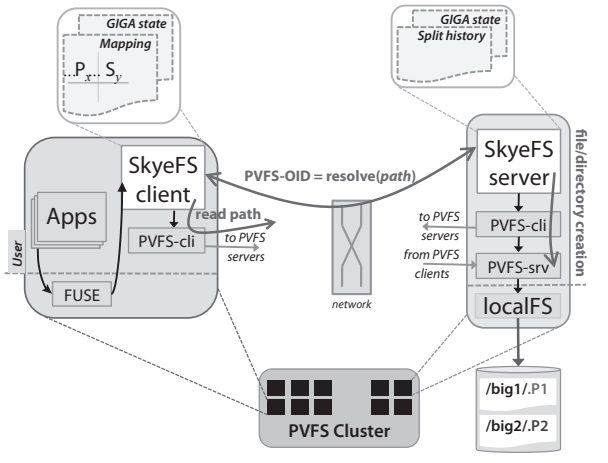
Shingled Magnetic Recording for Big Data Applications

Suresh, Gibson & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-12-105. May 2012.

Modern Hard Disk Drives (HDDs) are fast approaching the superparamagnetic limit forcing the storage industry to look for innovative ways to transition from traditional magnetic recording to Heat-Assisted Magnetic Recording or Bit-Patterned Magnetic Recording. Shingled Magnetic Recording (SMR) is a step in this direction as it delivers high storage capacity with minimal changes to current production infrastructure. However, since it sacrifices random-write capabilities of the device, SMR cannot be used as a drop-in replacement for traditional HDDs.

We identify two techniques to implement SMR. The first involves the insertion of a shim layer between the SMR device and the host, similar to the Flash Translation Layer found in Solid-State Drives (SSDs). The second technique, which we feel is the right direction for SMR, is to push enough intelligence up into the file system to effectively mask the sequential-write nature of the underlying SMR device. We present a custom-built SMR Device Emulator and ShingledFS, a FUSE-based SMR-aware file system that operates in tandem with the SMR Device Emulator. Our evaluation studies SMR for Big Data applications and we also examine the overheads introduced by the emulation. We show that Big Data workloads can be run effectively on SMR devices with an overhead as low as 2.2% after eliminating the overheads of emulation. Finally we present insights on garbage collection mechanisms and policies that will aid future SMR research.



SkyeFS Architecture.

Near-Real-Time Inference of File-Level Mutations from Virtual Disk Writes

Richter, Satyanarayanan, Harkes & Gilbert

Carnegie Mellon University School of Computer Science Technical Report CMU-CS-12-103. February 2012.

We describe a new mechanism for cloud computing enabling near-real-time monitoring of virtual disk write streams across an entire cloud. Our solution has low IO overhead for the guest VM, low latency to file-level mutation notification, and a layered design for scalability. We achieve low IO overhead by duplicating the virtual disk write stream as it passes through a managing VMM. We achieve low latency by performing semantic inference at as high a level as possible–file-level. We achieve cloud scale by layering our design allowing filtering of file-level mutations by each layer such that network traffic to centralized monitoring infrastructure is minimized. We assume this technique is used on pre-indexed virtual disks, most likely derived from a cooperating VM image library such as those used in clouds today. Our new cloud primitive enables system administration tasks that involve monitoring files–virus scanning, log file parsing, etc.–to be performed outside of the running VM instance, either on the VMM host, or shipped to a central monitoring agent.