PDL Packet Fall Update

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

PDL CONSORTIUM MEMBERS

American Power Conversion
EMC Corporation
Facebook
Fusion-io
Google
Hewlett-Packard Labs
Hitachi
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.

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/

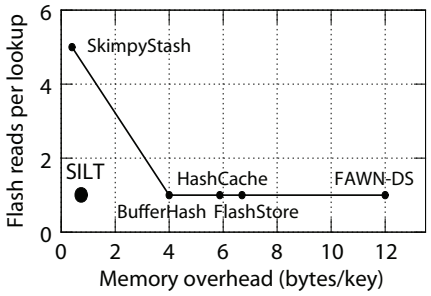
SELECTED RECENT PUBLICATIONS

SILT: A Memory-Efficient, High-Performance Key-Value Store

Lim, Fan, Andersen & Kaminsky

ACM Symposium on Operating Systems Principles (SOSP'11), Cascais, Portugal, October 2011.

SILT (Small Index Large Table) is a memory-efficient, high-performance key-value store system based on flash storage that scales to serve billions of key-value items on a single node. It requires only 0.7 bytes of DRAM per entry and retrieves key/value pairs using on average 1.01 flash reads each. SILT combines new algorithmic and systems techniques to balance the use of memory, storage, and computation. Our contributions include: (1) the design of three basic key-value stores each with a different emphasis on memory-efficiency and write-friendliness; (2) synthesis of the basic key-value stores to build a SILT key-value store system; and (3) an analytical model for tuning system parameters carefully to meet the needs of different workloads. SILT requires one to two orders of magnitude less memory to provide comparable throughput to current high-performance key-value systems on a commodity desktop system with flash storage.



The memory overhead and lookup performance of SILT and the recent key-value stores. For both axes, smaller is better.

YCSB++: Benchmarking and Performance Debugging Advanced Features in Scalable Table Stores

Patil, Polte, Ren, Tantisiriroj, Xiao, Lopez, Gibson, Fuchs & Rinaldi

Proc. of the 2nd ACM Symposium on Cloud Computing (SOCC '11), October 27–28, 2011, Cascais, Portugal.

Inspired by Google's BigTable, a variety of scalable, semistructured, weak-semantic table stores have been developed and optimized for different priorities such as query speed, ingest speed, availability, and interactivity. As these systems mature, performance benchmarking will advance from measuring the rate of simple workloads to understanding and debugging the performance of advanced features such as ingest speed-up techniques and function shipping filters from client to servers. This paper describes YCSB++, a set of extensions to the Yahoo! Cloud Serving Benchmark (YCSB) to improve performance understanding and debugging of these advanced features. YCSB++ includes multi- tester coordination for increased load and eventual consistency measurement, multi-phase workloads to quantify the consequences of work deferment and the benefits of anticipatory configuration optimization such as B-tree pre-splitting or bulk loading, and

continued on page 3

October 2011

Garth's 1988 RAID Paper Enters Hall of Fame

We are very pleased to announce that Garth Gibson's original RAID paper from SIGMOD 1988 — "A Case for Redundant Array of Inexpensive Disks" by Patterson, Gibson and Katz — was one of the four papers to be honored as a 2011 SIGOPS Hall of Fame Award paper. The award was made at the 23rd ACM Symposium on Operating Systems Principles (SOSP), October 23-26, 2011, Cascais, Portugal.

The SIGOPS Hall of Fame Award was instituted in 2005 to recognize the most influential Operating Systems papers that were published at least ten years in the past. The Hall of Fame Award Committee consists of past program chairs from SOSP, OSDI, EuroSys, past Weiser and Turing Award winners from the SIGOPS community, and representatives of each of the Hall of Fame Award papers.

August 2011

Intel Labs Invests $30M in the Future of Cloud and Embedded Computing with the Opening of Latest Intel Science and Technology Centers

Aimed at shaping the future of cloud computing and how increasing numbers of everyday devices will add computing capabilities, Intel Labs announced the latest Intel Science and Technology Centers (ISTC) for Cloud Computing Research (led by Greg Ganger, CMU and Phil Gibbons, Intel) and for Embedded Computing (led by Priya Narasimhan, CMU and Mei Chen, Intel), both headquartered at Carnegie Mellon University.

The ISTC for Cloud Computing forms a new cloud computing research community that broadens Intel's "Cloud 2015" vision with new ideas from top academic researchers, and includes research that extends and improves on Intel's existing cloud computing initiatives. The center combines top researchers from Carnegie Mellon University, Georgia Institute of Technology, University of California Berkeley, Princeton University, and Intel. The researchers will explore technology that will have has important future implications for the cloud, including built-in application optimization, more efficient and effective support of big data analytics on massive amounts of online data, and making the cloud more distributed and localized by extending cloud capabilities to the network edge and even to client devices.

In the future, these capabilities could enable a digital personal handler via a device wired into your glasses that sees what you see, to constantly pull data from the cloud and whisper information to you during the day — telling you who people are, where to buy an item you just saw, or how to adjust your plans when something new comes up.

Tapping into the expertise of leading researchers from Carnegie Mellon University, Cornell University, University of Illinois at Urbana Champaign, University of Pennsylvania, Pennsylvania State University, Georgia Institute of Technology, the University of California at Berkeley and Intel, the ISTC for embedded computing forms a new collaborative community to drive research to transform experiences in the home, car and retail environment of the future. With the growing popularity of mobile real-time and personalized technology, there is a corresponding rise in demand for specialized embedded computing systems to support a broad range of new applications — including many not yet envisioned.

A key area of research is to make it easier for these everyday devices to continuously collect, analyze and act on useful data from both sensors and online databases in a way that is timely, scalable and reliable. For example, in cars, this data could be used to customize in-vehicle entertainment options when specific passengers are recognized, and provide them better routing, retail, dining, and entertainment recommendations while on-the-road.

-- from the Intel News Room, by Connie Brown

June 2011

Onur Mutlu wins IEEE Young Computer Architect Award

ECE Assistant Professor Onur Mutlu has earned the inaugural IEEE Computer Society Technical Committee on Computer Architecture's Young Computer Architect Award "in recognition of outstanding contributions in the field of computer architecture in both research and education." The award recognizes outstanding contributions in the field of computer architecture by an individual who received their Ph.D. within six years of their nomination.

-- 8.5x11 News, June 23, 2011, Vol. 21, No. 49

continued on page 3

continued from page 2

June 2011

Satya Receives Outstanding Contributions Award at Mobisys'11

Congratulations to Prof. M. Satyanarayanan (Satya), who was awarded the SIGMOBILE 2010 Outstanding Contributions Award "for pioneering a wide spectrum of technologies in support of disconnected and weakly connected mobile clients" at Mobisys 2011. He joins an illustrious group of previous winners, including Prof. Daniel P. Siewiorek in 2006, who received the award "for pioneering fundamental contributions to wearable and context-aware computing." The SIGMOBILE Outstanding Contribution Award is given for significant and lasting contributions to the research on mobile computing and communications and wireless networking.

June 2011

PDL Alums win Best Demonstration at SIGMOD 2011

The demonstration of the DORA system ("A Data-oriented Transaction Execution Engine and Supporting Tools") won the Best Demonstration Award at SIGMOD 2011! The team that implemented the demo consisted of Ippokratis Pandis, Pinar Tozun, Miguel Branco, Dimitris Karampinas, Danica Porobic, Ryan Johnson and Natassa Ailamaki. The entire team is now affiliated with EPFL, with Ippokratis, Ryan and Natassa all recent members of the PDL. SIGMOD is the premier conference on data management systems, this year held in Athens, Greece.

June 2011

FAWN Team Winner of 2011 10GB JouleSort Daytona and Indy

The FAWN team, a joint Intel-CMU group, including Padmanabhan Pillai, Michael Kaminsky, Michael A. Kozuch, Vijay Vasudevan, Lawrence Tan and David G. Andersen won the 2011 10GB JouleSort competition using a Sandy Bridge-based platform with Intel SSDs. For more details see FAWNSort: Energy-efficient Sorting of 10GB and the Sort Benchmark home page.

June 2011

Swapnil Patil Receives ACM Student Research Award!

Swapnil Patil, a PhD student in computer science, took first place in the graduate student category of the Association for Computing Machinery (ACM) Student Research Competition Grand Finals. Patil received the award June 4 at the ACM Awards Banquet in San Jose, Calif. for his development of a file system director service that scales to millions of files, which he presented at SC10, the international conference for high performance computing, networking, storage and analysis. ACM's Student Research Program is sponsored by Microsoft Research to encourage students to pursue careers in computer science research, and to ensure the future of scientific discovery and innovation. The competitions, held at 13 major ACM Special Interest Group conferences within the last year, featured research projects produced by an international array of computer science graduate and undergraduate students. Winners from each of the SIG competitions were then eligible to compete in the Grand Finals.

continued from page 1

abstract APIs for explicit incorporation of advanced features in benchmark tests. To enhance performance debugging, we customized an existing cluster monitoring tool to gather the internal statistics of YCSB++, table stores, system services like HDFS, and operating systems, and to offer easy post-test correlation and reporting of performance behaviors. YCSB++ features are illustrated in case studies of two BigTable-like table stores, Apache HBase and Accumulo, developed to emphasize high ingest rates and fine-grained security.

ThermoCast: A Cyber-Physical Forecasting Model for Data Centers

Li, Liang, Liu, Nath, Terzis & Faloutsos

In KDD '11: Proceeding of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 21-24, 2011, San Diego, CA

Efficient thermal management is important in modern data centers as cooling consumes up to 50% of the total energy. Unlike previous work, we consider proactive thermal management, whereby servers can predict potential overheating events due to dynamics in data center configuration and workload, giving operators enough time to react. However, such forecasting is very challenging due to data center scales and complexity. Moreover, such a physical system is influenced by cyber effects, including workload scheduling in servers. We

continued on page 4

continued from page 3

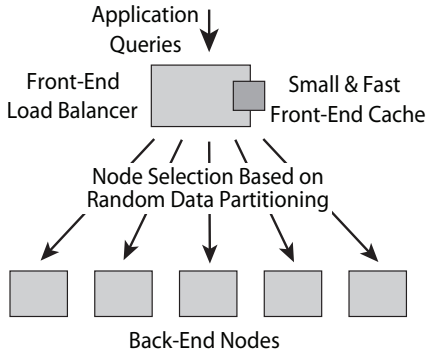
propose ThermoCast, a novel thermal forecasting model to predict the temperatures surrounding the servers in a data center, based on continuous streams of temperature and airflow measurements. Our approach is (a) capable of capturing cyberphysical interactions and automatically learning them from data; (b) computationally and physically scalable to data center scales; (c) able to provide online prediction with real-time sensor measurements. The paper's main contributions are: (i) We provide a systematic approach to integrate physical laws and sensor observations in a data center; (ii) We provide an algorithm that uses sensor data to learn the parameters of a data center's cyber-physical system. In turn, this ability enables us to reduce model complexity compared to full-fledged fluid dynamics models, while maintaining forecast accuracy; (iii) Unlike previous simulation-based studies, we perform experiments in a production data center. Using real data traces, we show that ThermoCast forecasts temperature 2 better than a machine learning approach solely driven by data, and can successfully predict thermal alarms 4.2 minutes ahead of time.

Small Cache, Big Effect: Provable Load Balancing for Randomly Partitioned Cluster Services

Fan, Lim, Andersen & Kaminsky

ACM Symposium on Cloud Computing (SOCC'11), Cascais, Portugal, October, 2011.

Load balancing requests across a cluster of back-end servers is critical for avoiding performance bottlenecks and meeting service-level objectives (SLOs) in large-scale cloud computing services. This paper shows how a small, fast popularity-based front-end cache can ensure load balancing for an important class of such services; furthermore, we prove an O(n log n) lower-bound on the necessary cache size and show that this size depends only on the total number of back-end nodes n, not the number of items stored in the system. We validate our analysis through simulation and empirical results running a key-value storage system on an 85-node cluster.



Small, fast cache at the front-end load balancer.

Reducing Memory Interference in Multicore Systems via Application-Aware Memory Channel Partitioning

Muralidhara, Subramanian, Mutlu, Kandemir & Moscibroda

Proceedings of the 44th International Symposium on Microarchitecture (MICRO), Porto Alegre, Brazil, December 2011.

Main memory is a major shared resource among cores in a multicore system. If the interference between different applications' memory requests is not controlled effectively, system performance can degrade significantly. Previous work aimed to mitigate the problem of interference between applications by changing the scheduling policy in the memory controller, i.e., by prioritizing memory requests from applications in a way that benefits system performance.

In this paper, we first present an alternative approach to reducing inter-application interference in the memory system: application-aware memory channel partitioning (MCP). The idea is to map the data of applications that are likely to severely interfere with each other to different memory channels. The key principles are to partition onto separate channels 1) the data of light (memory non-intensive) and heavy (memory-intensive) applications, 2) the data of applications with low and high row-buffer locality. Second, we observe that interference can be further reduced with a combination of memory channel partitioning and scheduling, which we call integrated memory partitioning and scheduling (IMPS). The key idea is to 1) always prioritize very light applications in the memory scheduler since such applications cause negligible interference to others, 2) use MCP to reduce interference among the remaining applications.

We evaluate MCP and IMPS on a variety of multi- programmed workloads and system configurations and compare them to four previously proposed state-of-the-art memory scheduling policies. Averaged over 240 workloads on a 24-core system with 4 memory channels, MCP improves system throughput by 7.1% over an application-unaware memory scheduler and 1% over the previous best scheduler, while avoiding modifications to existing memory schedulers. IMPS improves system throughput by 11.1% over an application- unaware scheduler and 5% over the previous best scheduler, while incurring much lower hardware complexity than the latter.

Memory Power Management via Dynamic Voltage/Frequency Scaling

David, Fallin, Gorbatov, Hanebutte & Mutlu

Proceedings of the 8th International Conference on Autonomic Computing (ICAC), Karlsruhe, Germany, June 2011.

Energy efficiency and energy-proportional computing have become a central focus in enterprise server architecture. As thermal and electrical

continued on page 5

continued from page 4

constraints limit system power, and datacenter operators become more conscious of energy costs, energy efficiency becomes important across the whole system. There are many proposals to scale energy at the datacenter and server level. However, one significant component of server power, the memory system, remains largely unaddressed.

We propose memory dynamic voltage/frequency scaling (DVFS) to address this problem, and evaluate a simple algorithm in a real system. As we show, in a typical server platform, memory consumes 19% of system power on average while running SPEC CPU2006 workloads. While increasing core counts demand more bandwidth and drive the memory frequency upward, many workloads require much less than peak bandwidth. These workloads suer minimal performance impact when memory frequency is reduced. When frequency reduces, voltage can be reduced as well.

We demonstrate a large opportunity for memory power reduction with a simple control algorithm that adjusts memory voltage and frequency based on memory bandwidth utilization. We evaluate memory DVFS in a real system, emulating reduced memory frequency by altering timing registers and using an analytical model to compute power reduction. With an average of 0.17% slowdown, we show 10.4% aver- age (20.5% max) memory power reduction, yielding 2.4% average (5.2% max) whole-system energy improvement.

Don't Settle for Eventual: Scalable Causal Consistency for Wide-Area Storage with COPS

Lloyd, Freedman, Kaminsky & Andersen

Proc. 23rd ACM Symposium on Operating Systems Principles (SOSP), Oct 2011.

Geo-replicated, distributed data stores that support complex online applications, such as social networks, must provide an "always on" experience where operations always complete with low latency. Today's systems often sacrifice strong consistency to achieve these goals, exposing inconsistencies to their clients and necessitating complex application logic. In this paper, we identify and define a consistency model—causal consistency with convergent conflict handling, or causal+—that is the strongest achieved under these constraints.

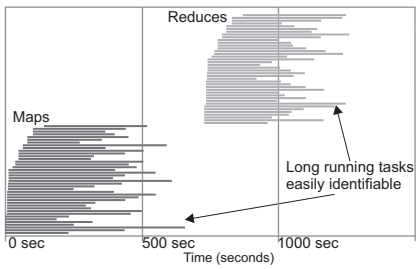
We present the design and implementation of COPS, a key-value store that delivers this consistency model across the wide-area. A key contribution of COPS is its scalability, which can enforce causal dependencies between keys stored across an entire cluster, rather than a single server like previous systems. The central approach in COPS is tracking and explicitly checking whether causal dependencies between keys are satisfied in the local cluster before exposing writes. Further, in COPS-GT, we introduce get transactions in order to obtain a consistent view of multiple keys without locking or blocking. Our evaluation shows that COPS completes operations in less than a millisecond, provides throughput similar to previous systems when using one server per cluster, and scales well as we increase the number of servers in each cluster. It also shows that COPS-GT provides similar latency, throughput, and scaling to COPS for common workloads.

Understanding and Improving the Diagnostic Workflow of MapReduce Users

Campbell, Ganesan, Gotow, Kavulya, Mulholland, Narasimhan, Ramasubramanian, Shuster & Tan

ACM Symposium on Computer Human Interaction for Management of Information Technology (CHIMIT), Boston, MA, December 2011.

New abstractions are simplifying the programming of large clusters, but



Swimlane graph charting the start and end times, and durations of Map and Reduce tasks for a single job. The graph also highlights the inherent structure of MapReduce jobs with map tasks completing before reduce tasks.

diagnosis nonetheless gets more and more challenging as cluster sizes grow: Debugging information increases linearly with cluster size, and the count of inter-component relationships grows quadratically. Worse, the new abstractions which simplified programming can also obscure the relationships between high-level (application) and low-level (task/process/disk/CPU) information flows. In this paper we analyze the workflow of several users and systems administrators connected with a large academic cluster (based the popular Hadoop implementation of the MapReduce abstraction) and propose improvements to the diagnosis-relevant information displays. We also offer a preliminary analysis of the efficacy of the changes we propose that demonstrates a 40% reduction in the time taken to accomplish 5 representative diagnostic tasks as compared to the current system.

Time Series Clustering: Complex is Simpler!

Li & Prakash

In Proceedings of the 28th International Conference on Machine Learning. June 28-July 2, 2011, Bellevue, WA.

Given a motion capture sequence, how to identify the category of the motion? Classifying human motions is a critical

continued on page 6

continued from page 5

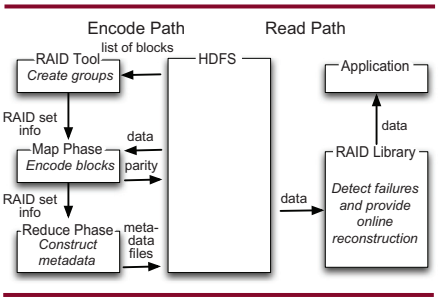
task in motion editing and synthesizing, for which manual labeling is clearly inefficient for large databases. Here we study the general problem of time series clustering. We propose a novel method of clustering time series that can (a) learn joint temporal dynamics in the data; (b) handle time lags; and (c) produce interpretable features. We achieve this by developing complex-valued linear dynamical systems (CLDS), which include real-valued Kalman filters as a special case; our advantage is that the transition matrix is simpler (just diagonal), and the transmission one easier to interpret. We then present Complex- Fit, a novel EM algorithm to learn the parameters for the general model and its special case for clustering. Our approach produces significant improvement in clustering quality, 1.5 to 5 times better than well-known competitors on real motion capture sequences.

DiskReduce: Replication as a Prelude to Erasure Coding in Data-Intensive Scalable Computing

Fan, Tantisiriroj, Xiao & Gibson

Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-11-112, October, 2011.

The first generation of Data-Intensive Scalable Computing file systems such as Google File System and Hadoop Distributed File System employed n replications for high data reliability, therefore delivering users only about 1/n of the total storage capacity of the raw disks. This paper presents DiskReduce, a framework integrating RAID into these replicated storage systems to significantly reduce the storage capacity overhead, for example, from 200% to 25% when triplicated data is dynamically replaced with RAID sets (e.g. 8 + 2 RAID 6 encoding). Based on traces collected from Yahoo!, Facebook and Opencloud cluster, we analyze (1) the capacity effectiveness of simple and not so simple strategies for grouping data



Encode and read path for RAID files.

blocks into RAID sets; (2) implication of reducing the number of data copies on read performance and how to overcome the degradation; and (3) different heuristics to mitigate "small write penalties." Finally, we introduce an implementation of our framework that has been built and submitted into the Apache Hadoop project.

Cyber-Physical-System Approach to Data Center Modeling and Control for Energy Efficiency

Parolini, Sinopoli, Krogh & Z. Wang

Proceedings of the IEEE, Special Issue on Cyber-Physical Systems, December 2011.

This paper presents data centers from a cyberphysical system (CPS) perspective. Current methods for controlling information technology (IT) and cooling technology (CT) in data centers are classified according to the degree to which they take into account both cyber and physical considerations. To evaluate the potential impact of coordinated CPS strategies at the data-center level, we introduce a control-oriented model that represents the data center as two coupled networks: a computational network representing the cyber dynamics and a thermal network representing the physical dynamics. These networks are coupled through the influence of the IT on both networks: servers affect both the quality of service (QoS) delivered by the computational network and the generation of heat in the thermal network. Using this model, three control strategies are evaluated with respect to their energy efficiency and computational performance: a baseline strategy that ignores CPS considerations, an uncoordinated strategy that manages the IT and CT independently, and a coordinated strategy that manages the IT and CT together to achieve optimal performance with respect to both QoS and energy efficiency. Simulation results show that the benefits to be realized from coordinating the control of IT and CT depend on the distribution and heterogeneity of the computational and cooling resources throughout the data center. A new cyber-physical index (CPI) is introduced as a measure of this combined distribution of cyber and physical effects in a given data center. We illustrate how the CPI indicates the potential impact of using coordinated CPS control strategies.

Switching the Optical Divide: Fundamental Challenges for Hybrid Electrical/Optical Datacenter Networks

Bazzaz, Tewari, Wang, Porter, Ng, Andersen, Kaminsky, Kozuch & Vahdat

Proc. 2nd ACM Symposium on Cloud Computing (SOCC), Oct 2011.

Recent proposals to build hybrid electrical (packet-switched) and optical (circuit switched) data center interconnects promise to reduce the cost, complexity, and energy requirements of very large data center networks. Supporting realistic traffic patterns, however, exposes a number of unexpected and difficult challenges to actually deploying these systems "in the wild." In this paper, we explore several of these challenges, uncovered during a year of experience using hybrid interconnects. We discuss both the problems that must be addressed to make these interconnects truly useful, and the implications of these challenges on what solutions are likely to be ultimately feasible.

continued on page 7

continued from page 6

WindMine: Fast and Effective Mining of Web-click Sequences

Sakurai, Li, Matsubara & Faloutsos

In 2011 Siam International Conference on Data Mining (SDM11). April 28-30, 2011, Mesa, AZ.

Given a large stream of users clicking on web sites, how can we find trends, patterns and anomalies? We have developed a novel method, WindMine, and its fine-tuning sibling, WindMine-part, to find patterns and anomalies in such datasets. Our approach has the following advantages: (a) it is effective in discovering meaningful "building blocks" and patterns such as the lunch-break trend and anomalies, (b) it automatically determines suitable window sizes, and (c) it is fast, with its wall clock time linear on the duration of sequences. Moreover, it can be made sub-quadratic on the number of sequences (WindMine-part), with little loss of accuracy.

We examine the effectiveness and scalability by performing experiments on 67 GB of real data (one billion clicks for 30 days). Our proposed WindMine does produce concise, informative and interesting patterns. We also show that WindMine-part can be easily implemented in a parallel or distributed setting, and that, even in a single-machine setting, it can be an order of magnitude faster (up to 70 times) than the plain version.

Draco: Top-Down Statistical Diagnosis of Large-scale VoIP Networks

Kavulya, Joshi, Hiltunen, Daniels, Gandhi & Narasimhan

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-11-109, April 2011.

Large scale integrated services such as VoIP running over IP networks are the future of telecommunications. The high availability requirements of such services require scalable techniques for rapid diagnosis and localization of user-visible failures. However, state-of-the-art network event correlation techniques often produce alarms that cannot easily be correlated to customer visible impacts because they work in a "bottom-up" fashion starting from device-level events and working upwards. In this paper, we develop a contrasting "top-down" approach to problem diagnosis that starts from user visible defects such as call drops and works downwards by identifying the network level elements that are the most suggestive of the defects. Our prototype, called Draco, uses statistical comparisons between good and bad system behavior to identify the underlying causes of problems without the need for any expert-provided rules or models, and without any prior training. This allows Draco to localize the causes of problems that have never been seen before. We have deployed Draco at scale for a portion of the VoIP operations of a major ISP. We demonstrate Draco's usefulness by provide examples of actual instances in which Draco helped operators diagnose service issues.

Practical Experiences with Chronics Discovery in Large Telecommunications Systems

Kavulya, Joshi, Hiltunen, Daniels, Gandhi & Narasimhan

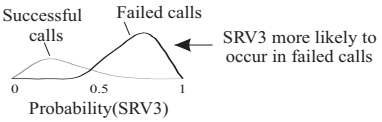
Workshop on System Logs and the Application of Machine Learning Techniques (SLAML), Cascais, Portugal, October 2011.

Chronics are recurrent problems that fly under the radar of operations teams because they do not perturb the system enough to set off alarms or violate service-level objectives. The discovery and diagnosis of never-before seen chronics poses new challenges as they are not detected by traditional threshold-based techniques, and many chronics can be present in a system at once, all starting and ending at different times. In this paper, we describe

1. Represent call attributes as truth table

SVR1	SVR2	SVR3	PHONE1	PHONE2	OUTCOME
1	1	0	0	0	SUCCESS
0	0	1	1	0	FAIL
0	0	1	0	1	FAIL

2. Model distribution of each attribute



An overview of steps used by our top-down, statistical diagnosis algorithm.

our experiences diagnosing chronics using server logs on a large telecommunications service. Our technique uses a scalable Bayesian distribution learner coupled with an information theoretic measure of distance (KL divergence), to identify the attributes that best distinguish failed calls from successful calls. Our preliminary results demonstrate the usefulness of our technique by providing examples of actual instances where we helped operators discover and diagnose chronics.

Six Degrees of Scientific Data: Reading Patterns for Extreme Scale Science IO

Lofstead, Polte, Gibson, Klasky, Schwan, Oldfield, Wolf & Liu

20th ACM Int. Symp. On High-Performance Parallel and Distributed Computing (HPDC'11), June 2011.

Petascale science simulations generate 10s of TBs of application data per day, much of it devoted to their checkpoint/restart fault tolerance mechanisms. Previous work demonstrated the importance of carefully managing such output to prevent application slowdown due to IO blocking, resource contention negatively impacting simulation performance and to fully exploit the IO bandwidth available to the petascale machine. This paper takes a further step in understanding and managing extreme-scale IO. Specifi-

continued on page 11

DISSERTATION ABSTRACT: Energy-efficient Data-intensive Computing with a Fast Array of Wimpy Nodes

Vijay Vasudevan

Carnegie Mellon University SCS
Ph.D. Dissertation, Oct. 10, 2011

Large-scale data-intensive computing systems have become a critical foundation for Internet-scale services. Their widespread growth during the past decade has raised datacenter energy demand and created an increasingly large financial burden and scaling challenge: Peak energy requirements today are a significant cost of provisioning and operating datacenters. In this thesis, we propose to reduce the peak energy consumption of datacenters by using a FAWN: A Fast Array of Wimpy Nodes. FAWN is an approach to building datacenter server clusters using low-cost, low-power servers that are individually optimized for energy efficiency rather than raw performance alone. FAWN systems, however, have a different set of resource constraints than traditional systems that can prevent existing software from reaping the improved energy efficiency benefits FAWN systems can provide.

This dissertation describes the principles behind FAWN and the software techniques necessary to unlock its energy efficiency potential. First, we present a deep study into building FAWN-KV, a distributed, log-structured key-value storage system designed for an early FAWN prototype. Second, we present a broader classification and workload analysis showing when FAWN can be more energy-efficient and under what workload conditions a FAWN cluster would perform poorly in comparison to a smaller number of high-speed systems. Last, we describe modern trends that portend a narrowing gap between CPU and I/O capability and highlight the challenges endemic to all future balanced systems. Using FAWN as an early example, we demonstrate that pervasive use of "vector interfaces" throughout distributed storage systems can improve throughput by an order of magnitude and eliminate the redundant work found in many data-intensive workloads.

Michelle Mazurek and Peter Klemperer discuss their work on Reactive Access Control at the 2011 PDL Spring Visit Day.

DISSERTATION ABSTRACT: Mining and Querying Multimedia Data

Fan Guo

Carnegie Mellon University SCS
Ph.D. Dissertation, Sept. 19, 2011

The emerging popularity of multimedia data, as digital representation of text, image, video and countless other milieus, with prodigious volumes and wild diversity, exhibits the phenomenal impact of modern technologies in reforming the way information is accessed, disseminated, digested and retained. This has iteratively ignited the data-driven perspective of research and development, to characterize perspicuous patterns, crystallize informative insights, and realize elevated experience for end-users, where innovations in a spectrum of areas of computer science, including databases, distributed systems, machine learning, vision, speech and natural languages, has been incessantly absorbed and integrated to elicit the extent and efficacy of contemporary and future multimedia applications and solutions.

Under the theme of pattern mining and similarity querying, this manuscript presents a number of pieces of research concerning multimedia data, to address an array of practical tasks encompassing automatic annotation, outlier detection, community discovery, multi-modal retrieval and learning to rank, in their respective contexts including satellite image analysis, internet traffic surveillance, image bioinformatics, and Web search. A repertoire of extant and novel techniques pertaining to graph mining, clustering analysis, tensor decomposition and probabilistic graphical models has been developed or adapted, which satisfactorily met differing quality and efficiency requisites postulated by specific application settings, best exemplified by the 40 times speed-up in annotating satellite images and the up to 30% performance improvement in predicting web search user clicks, yet without the loss of generality to similar and related scenarios.

DISSERTATION ABSTRACT: Performance Insulation: More Predictable Shared Storage

Matthew Wachs

Carnegie Mellon University SCS
Ph.D. Dissertation, Sept. 28, 2011

Many storage workloads do not need the performance afforded by a dedicated storage system, but do need the predictability and controllability that comes from one. Unfortunately, inter-workload interference, such as a reduction of locality when multiple request streams are interleaved, can result in dramatic loss of efficiency and performance.

Performance insulation is a system property where each workload sharing the system is assigned a fraction of resources (such as disk time) and receives nearly that fraction of its standalone

continued on page 9

continued from page 8

(dedicated system) performance. Because there is usually some overhead caused by sharing, there could be a drop in efficiency; but a system providing performance insulation provides a bound on efficiency loss at all times, called the R-value. We have built a storage server called Argon that achieves performance insulation in practice for R-values of 0.8-0.9. This means that, running together with other workloads on Argon, workloads lose, at most, only 10-20% of the efficiency they receive on a dedicated system.

While performance insulation provides a useful limit on loss of efficiency, many storage workloads also need performance guarantees. To ensure performance guarantees are consistently met, the appropriate allocation of resources needs to be determined and reserved, and later reevaluated if the workload changes in behavior or if the interference between workloads affects their ability to use resources effectively. If the resources assigned to a workload need to be increased to maintain its guarantee, but adequate resources are not available, violations will result.

Though intrinsic workload variability is fundamental, storage systems with the property of performance insulation strictly limit inter-workload interference, another source of variability. Such interference is the major source of "artificial" complexity in maintaining performance guarantees. We design and evaluate a storage system called Cesium that limits interference and thus avoids the class of guarantee violations arising from it. Workloads running on Cesium only suffer from those violations caused by their own variability and not those due to the activities of other workloads. Realistic and challenging workloads may experience an order of magnitude fewer violations running under Cesium. Performance insulation thus results in more reliable and efficient bandwidth guarantees.

DISSERTATION ABSTRACT: Fast Algorithms for Mining Co-evolving Time Series

Lei Li

Carnegie Mellon University SCS
Ph.D. Dissertation, Sept. 17, 2011

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

In this thesis, we focus on fast algorithms on mining co-evolving time series, with or without missing values. We will present a series of our effort in analyzing those data: (a) time series mining and summarization with missing values, and (b) learning features from multiple sequences. Algorithms proposed in the first work allow us to obtain meaningful patterns effectively and efficiently. Thus they enable vital mining tasks including forecast, compression, and segmentation for co-evolving time series, even with missing values. We also propose "PLiF" and Complex Linear Dynamical System (CLDS), novel algorithms to extract features from multiple sequences. Such features will serve as a corner stone of many applications for time series such clustering and similarity search. Our algorithms scale linearly with respect to the length of sequences, and outperform the competitors often by large factors. In addition, we will briefly mention several other time series mining problems and algorithms, including natural motion stitching, bone constrained occlusion filling, a parallelization of our algorithms for multi-core systems, and an forecasting algorithm for thermal conditions in data centers.

Alexey Tumanov, Ilari Shafer and Arkady Kanevsky at the 2011 PDL Spring Visit Day.

DISSERTATION ABSTRACT: Scalable Transaction Processing through Data-oriented Execution

Ippokratis Pandis

Carnegie Mellon University SCS
Ph.D. Dissertation, May 12, 2011

Data management technology changes the world we live in by providing efficient access to huge volumes of constantly changing data and by enabling sophisticated analysis of those data. While there has been an unprecedented increase in the demand for data management services; in parallel, we witness a tremendous shift in the underlying hardware toward highly parallel multicore processors. The data management systems in order to cope with the increased demand and user expectations, they need to exploit fully the abundantly available hardware parallelism. Transaction processing is one of the most important and challenging database workloads and this dissertation contributes in the quest for scalable transaction processing software. It shows that in the highly parallel multicore landscape the system designers should primarily focus on reducing the un-scalable critical sections of their systems, rather than improving the single-thread performance. In addition, it makes solid improvements in conventional transaction processing technology by avoiding executing unscalable critical sections in the lock

continued on page 10

continued from page 9

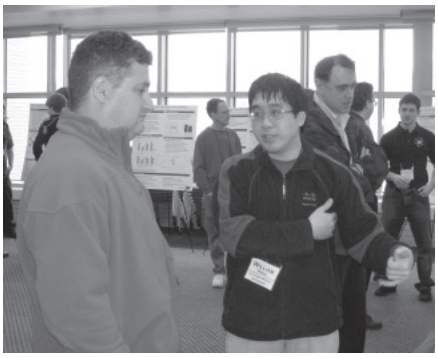
manager through caching, and in the log manager by downgrading them to composable ones. More importantly, it shows that conventional transaction processing has inherent scalability limitations due to the unpredictable access patterns caused by the request-oriented execution model it follows. Instead, it proposes to adopt a data-oriented execution model, and shows that transaction processing systems designed around data-oriented transaction execution break the inherent limitations of conventional execution. The data-oriented design paves the way for transaction processing systems to maintain scalability as parallelism increases for the foreseeable future; as hardware parallelism increases the benefits will only increase. In addition, the principles used to achieve scalability can generalize to other software systems facing similar scalability challenges with the shift to multicore hardware.

THESIS PROPOSAL:
Mining Tera-Scale Graphs with MapReduce: Theory, Engineering and Discoveries

U. Kang, SCS

October 20,2011

How do we find patterns and anomalies, on graphs with billions of nodes and edges, which do not fit in memory? How to use parallelism for such Tera- or Peta-scale graphs? In this thesis, we propose a carefully selected set of fundamental operations, that help answer those questions, including diameter estimation, solving eigenvalues, and inference on graphs. We package all these operations in PEGASUS, which, to the best of our knowledge, is the first such library, implemented on the top of the HADOOP platform, the open source version of MAPREDUCE. One of the key observations in this thesis is that many graph mining operations are essentially repeated matrix-vector multiplications. We describe a very important primitive for PEGASUS,



William Wang describes his research to Jeff Heller of NetApp.

called GIM-V (Generalized Iterated Matrix-Vector multiplication). GIM-V is highly optimized, achieving (a) good scale-up on the number of available machines, (b) linear running time on the number of edges, and (c) more than 9 times faster performance over the non-optimized version of GIM-V. Finally, we run experiments on real graphs. Our experiments ran on Disc-Cloud and M45, one of the largest HADOOP clusters available to academia. We report our findings on several real graphs, including one of the largest publicly available Web graphs, thanks to Yahoo! with ~6,7 billion edges. Some of our most impressive findings are (a) the discovery of adult advertisers in the who-follows-whom on Twitter, and (b) the 7-degrees of separation in the Web graph. Based on our current work, we propose the followings: large scale tensor analysis, graph layout for better compression, and anomaly detection in network data.

THESIS PROPOSAL:
Diagnosing Performance Changes by Comparing Request Flows

Raja Sambasivan, SCS

October 14, 2011

The causes of performance changes in a distributed system often elude even its developers. This proposed thesis develops a new technique for gaining insight into such changes: comparing request flows from two executions (e.g., of two system versions or time periods). Building on end-to-end request flow tracing within and across components, algorithms are described for identifying and ranking changes in the flow and/or timing of request processing. The implementation of these algorithms in a tool called Spectroscope is described and evaluated. Eight case studies are presented of using Spectroscope to diagnose performance changes in a prototype distributed storage service and in select Google services. To further show the generality of request-flow comparison, we also propose to adapt Spectroscope to work with HDFS and diagnose real problems observed within it.

MASTERS THESIS:
End-to-end Tracing in HDFS

William Wang , SCS

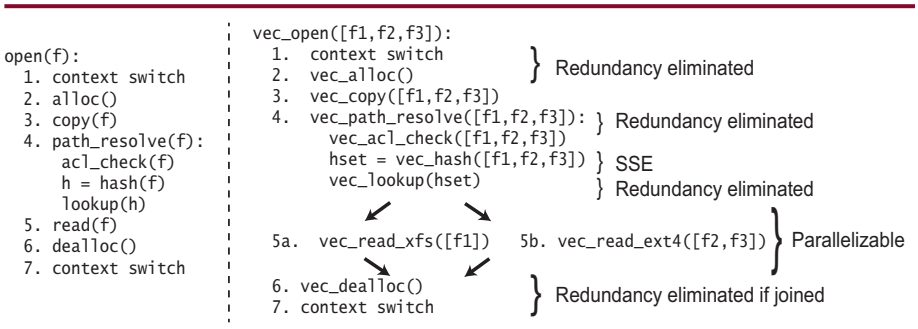
July 2011

Carnegie Mellon University School of Computer Science Technical Report CMU-CS-11-120, July 2011.

Debugging performance problems in distributed systems is difficult. Thus many debugging tools are being developed to aid diagnosis. Many of the most interesting new tools require information from end-to-end tracing in order to perform their analysis. This paper describes the development of an end-to-end tracing framework for the Hadoop Distributed File System. The approach to instrumentation in this implementation differs from previous ones as it focuses on detailed low-level instrumentation. Such instrumentation encounters the problems of large request flow graphs and a large number of different kinds of graphs, impeding the effectiveness of the diagnosis tools that use them. This report describes how to instrument at a fine granularity and explain techniques to handle the resulting challenges. The current implementation is evaluated in terms of performance, scalability, the data the instrumentation generates, and its ability to be used to solve performance problems.

continued from page 7

cally, its evaluations seek to understand how to efficiently read data for subsequent data analysis, visualization, check-point restart after a failure, and other read-intensive operations. In their entirety, these actions support the "end-to- end" needs of scientists enabling the scientific processes being undertaken. Contributions include the following. First, working with application scientists, we define 'read' benchmarks that capture the common read patterns used by analysis codes. Second, these read patterns are used to evaluate different IO techniques at scale to understand the effects of alternative data sizes and organizations in relation to the performance seen by end users. Third, defining the novel notion of a 'data district' to characterize how data is organized for reads, we experimentally compare the read performance seen with the ADIOS middleware's log-based BP format to that seen by the logically contiguous NetCDF or HDF5 formats commonly used by analysis tools. Measurements assess the performance seen across patterns and with different data sizes, organizations, and read process counts. Outcomes demonstrate that high end-to-end IO performance requires data organizations that offer flexibility in data layout and placement on parallel storage targets, including in ways that can make tradeoffs in the performance of data writes vs. reads.



Pseudocode for open() and proposed vec open(). vec open() provides opportunities for eliminating redundant code execution, vector execution when possible, and parallel execution otherwise.

The Case for VOS: The Vector Operating System

Vasudevan, Andersen & Kaminsky

In 13th Workshop on Hot Topics in Operating Systems (HotOS 2011). May 2011.

Operating systems research for many-core systems has recently focused its efforts on supporting the scalability of OS-intensive applications running on increasingly parallel hardware. Lost amidst the march towards this parallel future is efficiency: Perfectly parallel software may saturate the parallel capabilities of the host system, but in doing so can waste hardware resources. This paper describes our motivation for the Vector OS, a design inspired by vector processing systems that provides efficient parallelism. The Vector OS organizes and executes requests for operating system resources through "vector" interfaces that operate on vectors of objects. We argue that these interfaces allow the OS to capitalize on numerous chances to both eliminate redundant work found in OS-intensive systems and use the underlying parallel hardware to its full capability, opportunities that are missed by existing operating systems.

Failure Diagnosis of Complex Systems

Kavulya, Josh, Di Giandomenico & Narasimhan

To appear in "*Resilience Assessment and Evaluation.*" Springer Verlag, 2011.

Failure diagnosis is the process of identifying the causes of impairment in a system's function based on observable symptoms, i.e., determining which fault led to an observed failure. Since multiple faults can often lead to very similar symptoms, failure diagnosis is often the first line of defense when things go wrong — a prerequisite before any corrective actions can be undertaken. The results of diagnosis also provide data about a system's operational fault profile for use in offline resilience evaluation. While diagnosis has historically been a largely manual process requiring significant human input, techniques to automate as much of the process as possible have significantly grown in importance in many industries including telecommunications, internet services, automotive systems, and aerospace. This chapter presents a survey of automated failure diagnosis techniques including both model-based and model-free approaches. Industrial applications of these techniques in the above domains are presented, and finally, future trends and open challenges in the field are discussed.

Privacy-Sensitive VM Retrospection

Richter, Ammons, Harkes, Goode, Bila, De Lara, Bala & Satyanarayanan

HotCloud 2011 3rd USENIX Workshop on Hot Topics in Cloud Computing. Portland, OR, June 14-17, 2011.

The success of cloud computing leads to large, centralized collections of virtual machine (VM) images. The ability to retrospect (examine the historical state of) these images at a high semantic level can be valuable in many aspects of IT management such as debugging and troubleshooting, software quality control, legal establishment of data or code provenance, and cyber forensics such as malware tracking and licensing violations. In this paper, we explore

continued on page 12

RECENT PUBLICATIONS

continued from page 11

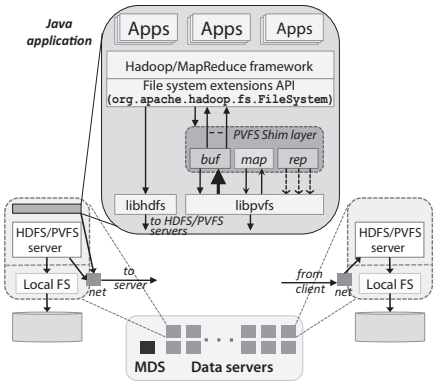
the privacy implications of VM retrospection. We argue that retrospection will worsen current concerns about privacy in cloud computing. We develop privacy-sensitive requirements for the design of a retrospection mechanism, and then show how they can be met in a functional prototype.

Diagnosis in Automotive Systems: A Survey

Lanigan, Kavulya, Narasimhan & Salman

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-11-110. June 2011.

Modern automotive electronic control systems are distributed, networked embedded systems. Diagnostic routines implemented on individual components cannot adequately identify the true cause of anomalous behavior because their view is restricted to component-local information. A growing trend in diagnostics research for these systems is to use system-level approaches to diagnose anomalous behavior and provide a consistent, global view of the system's health. Current approaches are typically motivated by a desire to improve either off-line maintenance or run-time safety.



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

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

Tantisiriroj, Patil, Gibson, Son, Lang & Ross

Supercomputing 2011, November 12-18, 2011, Seattle, Washington USA.

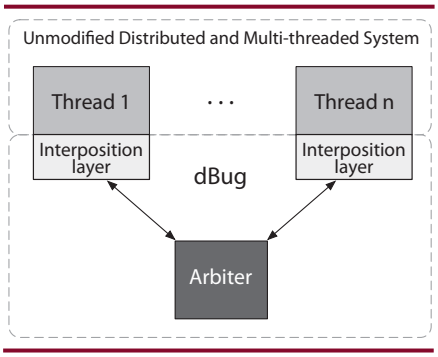
Data-intensive applications fall into two computing styles: Internet services (cloud computing) or high-performance computing (HPC). In both categories, the underlying file system is a key component for scalable application performance. In this paper, we explore the similarities and differences between PVFS, a parallel file system used in HPC at large scale, and HDFS, the primary storage system used in cloud computing with Hadoop. We integrate PVFS into Hadoop and compare its performance to HDFS using a set of data-intensive computing benchmarks. We study how HDFS-specific optimizations can be matched using PVFS and how consistency, durability, and persistence tradeoffs made by these file systems affect application performance. We show how to embed multiple replicas into a PVFS file, including a mapping with a complete copy local to the writing client, to emulate HDFS's file layout policies. We also highlight implementation issues with HDFS's dependence on disk bandwidth and benefits from pipelined replication.

dBug: Systematic Testing of Distributed and Multi-threaded Systems

Simsa, Bryant, Gibson

18th International Workshop on Model Checking of Software (SPIN'11), Snowbird UT, July 2011.

In order to improve quality of an implementation of a distributed and multi-threaded system, software engineers inspect code and run tests. However, the concurrent nature of such systems makes these tasks challenging.



dBug Architecture.

For testing, this problem is addressed by stress testing, which repeatedly executes a test hoping that eventually all possible outcomes of the test will be encountered. In this paper we present the dBug tool, which implements an alternative method to stress testing called systematic testing. The systematic testing method implemented by dBug controls the order in which certain concurrent function calls occur. By doing so, the method can systematically enumerate possible inter- leavings of function calls in an execution of a concurrent system. The dBug tool can be thought of as a light-weight model checker, which uses the implementation of a distributed and multi-threaded system and its test as an implicit description of the state space to be explored. In this state space, the dBug tool performs a reachability analysis checking for a number of safety properties including the absence of 1) deadlocks, 2) conflicting non-reentrant function calls, and 3) system aborts and runtime assertions inserted by the user.

Garth Gibson and Jerry Fredin discuss research at the PDL Spring Visit Day.