PDL Packet Spring Update

NEWSLETTER ON THE PARALLEL DATA LABORATORY • SPRING 2004

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

CONSORTIUM MEMBERS

EMC
Hewlett-Packard
Hitachi
Hitachi Global Storage Tech.
IBM
Intel
LSI Logic
Microsoft Research
Network Appliance
Panasas
Oracle
Seagate
Sun Microsystems
Veritas

CONTENTS

THE PDL PACKET

EDITOR
Joan Digney

CONTACT
Greg Ganger
PDL Director

Karen Lindenfelser
PDL Business Administrator

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

A Framework for Building Unobtrusive Disk Maintenance Applications

Thereska, Schindler, Bucy, Salmon, Lumb & Ganger

Proceedings of the 3rd USENIX Conference on File and Storage Technologies (FAST '04). San Francisco, CA. March 31, 2004. Best Student Paper Award.

This paper describes a programming model and system support for clean construction of disk maintenance applications. Such applications expose the disk activity to be done, and then process completed requests as they are reported. The system ensures that these applications make steady forward progress without competing for disk access with a system's primary applications. It opportunistically completes maintenance requests by using disk idle time and free-block scheduling. In this paper, three disk maintenance applications (backup, write-back cache destaging, and disk layout reorganization –see figure below) are adapted to the system support and evaluated on a FreeBSD implementation. All are shown to successfully execute in busy systems with minimal (e.g., <2%) impact on foreground disk performance. In fact, by modifying FreeBSD's cache to write dirty blocks for free, the average read cache miss response time is decreased by 15–30%.

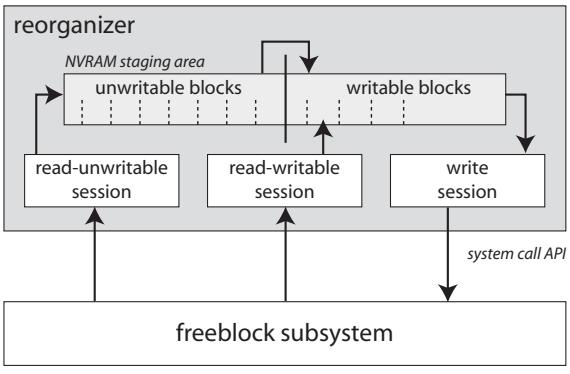
File Classification in Self-* Storage Systems

Mesnier, Thereska, Ellard, Ganger & Seltzer

Proceedings of the First International Conference on Autonomic Computing (ICAC-04). New York, NY. May 2004.

To tune and manage themselves, file and storage systems must understand key properties (e.g., access pattern, lifetime, size) of their various files. This paper describes how systems can automatically learn to classify the properties of files (e.g., read-only access pattern, short-lived, small in size) and predict the properties of new files, as they are created, by exploiting the strong associations between a file's properties and the names and attributes assigned to it. These associations exist, strongly but differently, in each of four real NFS environments studied. Decision tree classifiers can automatically identify and model such associations, providing prediction accuracies that often exceed 90%. Such predictions can be used to select storage policies (e.g., disk allocation schemes and replication factors) for individual files. Further, changes in associations can expose information about applications, helping autonomic system components distinguish growth from fundamental change.

… continued on pg. 2



Layout reorganizer architecture. This diagram illustrates the design of the layout reorganizer implemented using our framework. The read-unwritable session manages blocks whose dependencies have not yet been solved. The read-writable session manages all blocks that can be read because their dependencies have been solved. The write session manages all block writes. All data is temporarily stored in the NVRAM staging area.

PDL NEWS

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

April 2004
PDL Researchers Receive Best Student Paper Award at FAST '04

Congratulations to PDL researchers Eno Thereska, Jiri Schindler, John Bucy, Brandon Salmon and Gregory R. Ganger, who have been awarded Best Student Paper by the program committee of the USENIX Conference on File and Storage technologies (FAST '04) for their paper "A Framework for Building Unobtrusive Disk Maintenance Applications."

The FAST Best Student Paper award was also given to PDL researchers in 2002 when Jiri Schindler, John Linwood Griffin, Christopher R. Lumb, and Gregory R. Ganger were recognized for their paper "Track-Aligned Extents: Matching Access Patterns to Disk Drive Characteristics."

April 2004
Best Paper Award at ICDE 2004

Shimin Chen, Anastassia Ailamaki, Phillip Gibbons, and Todd Mowry received the best paper award for "Improving Hash Join Performance through Prefetching" at the International Conference on Data Engineering (ICDE) 2004. The conference took place in Boston, MA from March 30 through April 2. ICDE is one of the top database conferences, with hundreds of submitted papers, and extremely selective acceptance ratio (typically, 1-of-5 to 1-of-7). The paper focuses on the most expensive database operation ('join'), and proposes novel methods to accelerate it.

March 2004
CMU Sensor Detects Computer Hard Drive Failures

The Critter Temperature Sensor, a new heat-sensitive sensor to detect computer hard drive failures, which attaches to a user's desktop computer, is being deployed across campus to monitor the working environment of university computers, according to Michael Bigrigg, a project scientist for the PDL.

"We are trying save the life of the computer hard drive. Hard drives get hot and the sensor is designed to pick up

… continued on pg. 4

RECENT PUBLICATIONS

… continued from pg. 1

MEMS-based Storage Devices and Standard Disk Interfaces: A square peg in a round hole?

Schlosser & Ganger

Proceedings of the 3rd USENIX Conference on File and Storage Technologies (FAST '04). San Francisco, CA. March 31, 2004.

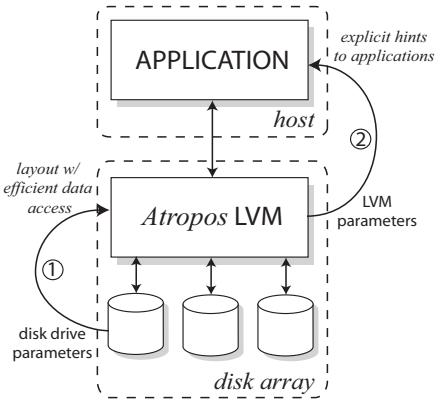
MEMS-based storage devices are a new technology that is significantly different from both disk drives and semiconductor memories. These differences motivate the question of whether they need new abstractions to be utilized by systems, or if existing abstractions will work well. This paper addresses this question by examining the fundamental reasons that the abstraction works for existing systems, and by showing that these reasons hold for MEMS-based storage. This result is borne out through several case studies of proposed roles MEMS-based storage devices may take in future systems, and potential policies that may be used to tailor systems' access to MEMS-based storage. We argue that when considering the use of MEMS-based storage in systems, their performance should be compared to that of a hypothetical disk drive that matches the speed of a MEMS-based storage device. We discuss exceptional workloads that can use specific features of MEMS-based storage devices and that may require extensions to current abstractions. Also, we consider the ramifications of the assumptions that are made in today's models of MEMS-based storage devices.

Atropos: A Disk Array Volume Manager for Orchestrated Use of Disks

Schindler, Schlosser, Shao, Ailamaki & Ganger

Proceedings of the 3rd USENIX Conference on File and Storage Technologies (FAST '04). San Francisco, CA. March 31, 2004.

The Atropos logical volume manager allows applications to exploit characteristics of its underlying collection of disks. It stripes data in track-sized units and explicitly exposes the boundaries, allowing applications to maximize efficiency for sequential access patterns even when they share the array. Further, it supports efficient diagonal access to blocks on adjacent tracks, allowing applications to orchestrate the layout and access of two-dimensional data structures, such as relational database tables, to maximize performance for both row-based and column-based accesses.



Atropos logical volume manager architecture. Atropos exploits disk characteristics (arrow 1), automatically extracted from disk drives, to construct a new data organization. It exposes high-level parameters that allow applications to directly take advantage of this data organization for efficient access to one- or two-dimensional data structures (arrow 2).

… continued on pg. 3

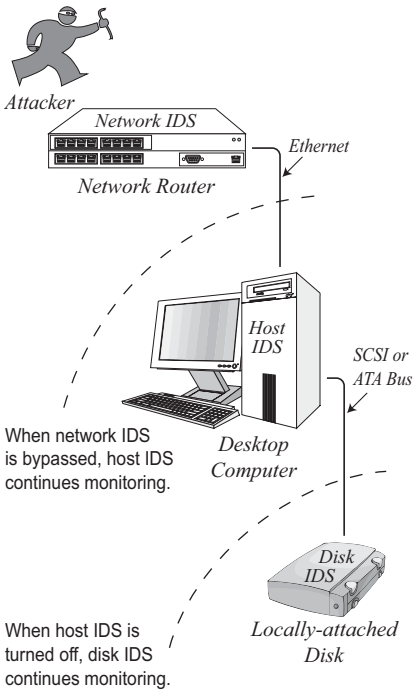
… continued from pg. 2

On the Feasibility of Intrusion Detection Inside Workstation Disks

Griffin, Pennington, Bucy, Choundappan, Muralidharan & Ganger

Carnegie Mellon Parallel Data Lab Technical Report CMU-PDL-03-106, December 2003.

Storage-based intrusion detection systems (IDSes) can be valuable tools in monitoring for and notifying administrators of malicious software executing on a host computer, including many common intrusion toolkits. This paper makes a case for implementing IDS functionality in the firmware of workstations' locally attached disks, on which the bulk of important system files typically reside. To evaluate the feasibility of this approach, we built a prototype disk-based IDS into a SCSI disk emulator. Experimental results from this prototype indicate that it would indeed be feasible, in terms of CPU and memory costs, to include IDS functionality in low-cost desktop disk drives.



The role of a disk-based intrusion detection system (IDS). A disk-based IDS watches over all data and executable files that are persistently written to local storage, monitoring for suspicious activity that might indicate an intrusion on the host computer.

Integrating Portable and Distributed Storage

Tolia, Harkes, Kozuch & Satyanarayanan

Proceedings of the 3rd USENIX Conference on File and Storage Technologies (FAST '04). San Francisco, CA. March 31, 2004.

We describe a technique called lookaside caching that combines the strengths of distributed file systems and portable storage devices, while negating their weaknesses. In spite of its simplicity, this technique proves to be powerful and versatile. By unifying distributed storage and portable storage into a single abstraction, lookaside caching allows users to treat devices they carry as merely performance and availability assists for distant file servers. Careless use of portable storage has no catastrophic consequences. Experimental results show that significant performance improvements are possible even in the presence of stale data on the portable device.

Dynamic Quarantine of Internet Worms

Wong, Wang, Song, Bielski & Ganger

The International Conference on Dependable Systems and Networks (DSN 2004). Palazzo dei Congressi, Florence, Italy. June 28th to July 1, 2004.

If we limit the contact rate of worm traffic, can we alleviate and ultimately contain Internet worms? This paper sets out to answer this question. Specifically, we are interested in analyzing different deployment strategies of rate control mechanisms and the effect thereof on suppressing the spread of worm code. We use both analytical models and simulation experiments. We find that rate control at individual hosts or edge routers yields a slowdown that is linear in the number of hosts (or routers) with the rate limiting filters. Limiting contact rate at the backbone routers, however, is substantially more effective—it renders a slowdown comparable to deploying rate limiting filters at every individual host. This result holds true even when susceptible and infected hosts are patched and immunized dynamically. To provide context for the analysis, we examine real traffic traces obtained from a campus of computing network. We observe that rate throttling could be enforced with minimal impact on legitimate communications. Two worms observed in the traces, however, would be significantly slowed.

A Protocol Family for Versatile Survivable Storage Infrastructures

Goodson, Wylie, Ganger & Reiter

The International Conference on Dependable Systems and Networks (DSN 2004). Palazzo dei Congressi, Florence, Italy. June 28th to July 1, 2004.

Survivable storage systems mask faults. A protocol family shifts the decision of which types of faults from implementation time to data-item creation time. If desired, each data-item can be protected from different types and numbers of faults. This paper describes and evaluates a family of storage access protocols that exploit data versioning to efficiently provide consistency for erasure-coded data. This protocol family supports a wide range of fault models with no changes to the client-server interface or server implementations. Its members also shift overheads to clients. Readers only pay these overheads when they actually observe concurrency or failures. Measurements of a prototype block-store show the efficiency and scalability of protocol family members.

Improving Hash Join Performance through Prefetching

Chen, Ailamaki, Gibbons & Mowry

Proceedings of the 20th International Conference on Data Engineering (ICDE 2004). Boston, MA. March 30 to April 2, 2004. Best Paper Award.

Hash join algorithms suffer from extensive CPU cache stalls. This paper shows that the standard hash join algorithm for disk-oriented databases (i.e.

… continued on pg. 4

RECENT PUBLICATIONS

… continued from pg. 3

GRACE) spends over 73% of its user time stalled on CPU cache misses, and explores the use of prefetching to improve its cache performance. Applying prefetching to hash joins is complicated by the data dependencies, multiple code paths, and inherent randomness of hashing. We present two techniques, group prefetching and software-pipelined prefetching, that overcome these complications. These schemes achieve 2.0–2.9X speedups for the join phase and 1.4–2.6X speedups for the partition phase over GRACE and simple prefetching approaches. Compared with previous cache-aware approaches (i.e. cache partitioning), the schemes are at least 50% faster on large relations and do not require exclusive use of the CPU cache to be effective.

Paul Massiglia and John Griffin discuss storage in Colorado.

PDL NEWS

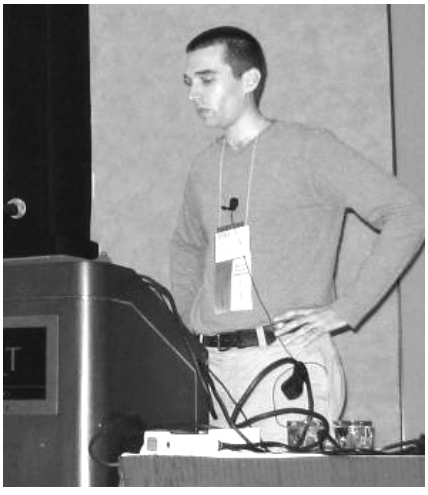
… continued from pg. 2

the slightest temperature variation," Bigrigg said. He added that the sensor will help researchers understand wasted energy with the hope of extending the lifespan of a computer hard drive by sensing how much daily heat a hard drive endures. So far, the new sensor, the size of a dime, has been deployed in offices and labs throughout Carnegie Mellon's Hamburg Hall.

—ece news & events, March 1, 2004

February 2004
Network Appliance Donates Filer for PDL Storage Needs

Network Appliance has donated a FAS900 series filer with 2 Terabytes of raw capacity and all of the software bells and whistles, with a retail value of $170K, in all. This filer will be used for critical PDL storage needs, including a software development repository, the PDL web server, and Lab member home directories.



Steve Schlosser presents his talk on MEMS-Based Storage at FAST 2004.

February 2004
Mowry to Head Intel Lab

Todd Mowry, associate professor of computer science, will succeed Mahadev Satyanarayanan as head of Intel Research Pittsburgh, effective this May. Mowry will bring a new research thrust to the lablet at the intersection of databases, architecture, compilers and operating systems. According to Satyanarayanan, in the two short years of its existence, Intel Research Pittsburgh is already making a big impact on a number of areas of research, including personal computing mobility (Internet Suspend/Resume project), wide-area sensing (IrisNet project), and interactive search of complex data (Diamond project). "We are clearly past the start-up phase, and can look forward to continued growth and many more accomplishments in 2004 and beyond," Satyanarayanan said.

—cmu.misc.news, Feb. 3, 2004

January 2004
Seagate supports the Self-* Storage Project with Equipment Donation

Greg Ganger (Assoc. Professor, ECE and CS) and the Parallel Data Lab (PDL) have received a $25K equipment grant of high-end SCSI disks from Seagate. The grant significantly increases the capacity of the testbed for PDL's new Self-* Storage project, which seeks to create large-scale self-managing, self-organizing, self-tuning storage systems from generic servers.

January 2004
Spiros Papadimitriou wins a Best Paper Award at VLDB03

Computer Science Ph.D. candidate Spiros Papadimitriou has received a Best Paper Award from the Very Large Data Bases (VLDB) 2003 Conference for his paper "Adaptive, Hands-Off Stream Mining," which was co-authored with Anthony Brockwell and Christos Faloutsos. The conference took place this past September in Berlin and is one of the most prestigious and selective database conferences. The paper is available on our publications page.

— with info, CMU 8.5 x 11 News, Jan. 8, 2004

January 2004
LSI Logic Joins PDL Industrial Research Consortium

The PDL is pleased to announce that LSI Logic has joined the PDL Consortium of companies that support and participate in PDL research. From the LSI Logic page: Founded in 1981, LSI Logic pioneered the ASIC (Application Specific Integrated Circuit) industry. LSI Logic is a leading designer and manufacturer of communications, consumer and storage semiconductors for applications that access, interconnect and store data, voice and video. Today, the company focuses on providing highly complex ASICs, ASSPs (Application Specific Standard Products), RapidChip™, host bus adapters, software and storage systems.