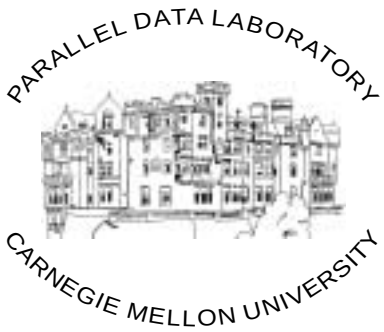

THE
PDL Packet Spring Update

THE NEWSLETTER ON PARALLEL DATA SYSTEMS • SPRING 2000

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

CONSORTIUM MEMBERS

CLARiiON Array Development
EMC Corporation
Hewlett-Packard Laboratories
Hitachi, Limited
Infineon Technologies
Intel Corporation
LSI Logic
MTI Technology Corporation
Novell Corporation
PANASAS, L.L.C.
ProCom Technology
Quantum Corporation
Seagate Technology
Sun Microsystems
Veritas Software Corporation
3Com Corporation

I◆N◆S◆I◆D◆E

THE PDL PACKET

PUBLISHER
David Nagle, PDL Director

EDITOR
Joan Digney

CONTACT
Patty Mackiewicz
PDL Business Administrator

The Parallel Data Laboratory
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
VOICE 412•268•6716
FAX 412•268•3010

RECENT PUBLICATIONS: ABSTRACTS

http://www.pdl.cs.cmu.edu/publications/publications.html

Data Mining on an OLTP System (Nearly) for Free

Riedel, Faloutsos, Ganger & Nagle

Proceedings of ACM SIGMOD 2000 International Conference on Management of Data, Dallas, TX, May 14-19.

This paper proposes a scheme for scheduling disk requests that takes advantage of the ability of high-level functions to operate directly at individual disk drives. We show that such a scheme makes it possible to support a Data Mining workload on an OLTP system almost for free: there is only a small impact on the throughput and response time of the existing workload. Specifically, we show that an OLTP system has the disk resources to consistently provide one third of its sequential bandwidth to a background Data Mining task with close to zero impact on OLTP throughput and response time at high transaction loads. At low transaction loads, we show much lower impact than observed in previous work.This means that a production OLTP system can be used for Data Mining tasks without the expense of a second dedicated system. Our scheme takes advantage of close interaction with the on-disk scheduler by reading blocks for the Data Mining workload as the disk head "passes over" them while satisfying demand blocks from the OLTP request stream. We show that this scheme provides a consistent level of throughput for the background workload even at very high foreground loads. Such a scheme is of most benefit in combination with an Active Disk environment that allows the background Data Mining application to also take advantage of the processing power and memory available directly on the disk drives.

Highly Concurrent Shared Storage

Amiri, Gibson & Golding

Proceedings of the International Conference on Distributed Computing Systems, Taipei, April 2000.

Switched system-area networks enable thousands of storage devices to be shared and directly accessed by end hosts, promising databases and filesystems highly scalable, reliable storage. In such systems, hosts perform access tasks (read and write) and management tasks (storage migration and reconstruction of data on failed devices.) Each task translates into multiple phases of low-level device I/Os, so that concurrent host tasks accessing shared devices can corrupt redundancy codes and cause hosts to read inconsistent data. Concurrency control protocols that scale to large system sizes are required in order to coordinate on-line storage management and access tasks. In this paper, we identify the tasks that storage controllers must perform, and propose an approach which allows these tasks to be composed from basic operations-called base storage transactions (BSTs)-such that correctness requires only the serializability of the BSTs and not of the parent tasks. We present highly scalable distributed protocols which exploit storage technology trends and BST properties to achieve serializability while coming within a few percent of ideal performance.

... continued on pg. 2

PDL NEWS

February 2000

Hui Zhang Selected Sloan Foundation Fellow

Hui Zhang, Finmeccanica Assistant Professor School of Computer Science and Department of Electrical and Computer Engineering, has been selected as a Sloan Foundation Fellow. This is a highly competitive program for junior faculty in six fields: chemistry, computer science, economics, mathematics, neuroscience, and physics with only 100 fellowships awarded per year. The fellowship provides a $40,000 grant over a two-year period. Dr Zhang's research interests focus on scalable solutions for Quality of Service and value-added distributed services over the Internet. He is involved in several projects including Darwin, Libra, Gemini, Indra, and an NSF Career Award Project.

March 25, 2000

Tim, the Enchanter (a.k.a. The Rock) Arrives

No qustion about it – Timothy is a Ganger. In true Ganger male fashion, he barely noted the passing of the official deadline, asked to know the final extended deadline (Sunday's induction), and slipped in just under it (8:57p.m. on Saturday, March 25). Timothy's got his mother's face, his father's ability to let people know when his drawers are messy, and Shaq-like size. At 22 inches and 9 pounds, 12 ounces, he's not going to let the other babies on the block push him around! Congratulations Greg and Jenny!

Greg and Jenny Ganger welcome their son Timothy on March 25!

RECENT PUBLICATIONS

... continued from pg. 1

Modeling and Performance of MEMS-Based Storage Devices

Griffin, Schlosser, Ganger & Nagle

Proceedings of ACM SIGMETRICS 2000, Santa Clara, CA, June 17-21.

MEMS-based storage devices are seen by many as promising alternatives to disk drives. Fabricated using conventional CMOS processes, MEMS-based storage consists of thousands of small, mechanical probe tips that access gigabytes of high-density, nonvolatile magnetic storage. This paper takes a first step towards understanding the performance characteristics of these devices by mapping them onto a disk-like metaphor. Using simulation models based on the mechanics equations governing device operation, this work explores how different physical characteristics (e.g., actuator forces and per-tip data rates) impact the design trade-offs and performance of MEMS-based storage. Overall results indicate that average access times for MEMS-based storage are 6.5 times faster than for a modern disk (1.5 ms vs. 9.7 ms). Results from filesystem and database benchmarks show that this improvement reduces application I/O stall times up to 70%, resulting in overall performance improvements of 3X.

Towards Higher Disk Head Utilization: Extracting "Free" Bandwidth From Busy Disk Drives

Lumb, Schindler, Ganger, Riedel & Nagle

CMU SCS Technical Report, CMU-CS-00-130, May 2000.

Freeblock scheduling is a new approach to utilizing more of disks' potential media bandwidths. By filling rotational latency periods with useful media transfers, 20-50% of a never-idle disk's bandwidth can be provided to background applications with no effect on foreground response times. This paper describes freeblock scheduling and demonstrates its value with two concrete applications: free segment cleaning and free data mining. Free segment cleaning allows an LFS file system to maintain its ideal write performance when cleaning overheads would otherwise cause up to factor of 3 performance decreases. Free data mining can achieve 45-70 full disk scans per day on an active transaction processing system, with no effect on transaction performance.

Operating System Management of MEMS-based Storage Devices

Griffin, Schlosser, Ganger & Nagle

CMU SCS Technical Report, CMU-CS-00-136, May 2000.

MEMS-based storage devices promise significant performance, reliability and power improvements relative to disk drives. This paper explores how the physical characteristics of these devices change four aspects of operating system management: request scheduling, data placement, failure management and power management. Disk request scheduling algorithms are adapted and found to

... continued on pg. 3

RECENT PUBLICATIONS

... continued from pg. 2

be appropriate for these devices. However, new data placement schemes are shown to better match their differing mechanical positioning characteristics. With aggressive internal redundancy, MEMS-based storage devices can tolerate failure modes that cause disk data loss. As well, MEMS-based storage devices enable a finer granularity of OS-level power management as the devices can be stopped and started rapidly and their mechanical components can be individually enabled or disabled to reduce power consumption.

Designing Computer Systems with MEMS-Based Storage

Schlosser, Griffin, Nagle & Ganger

CMU SCS Technical Report, CMU-CS-00-137, May 2000.

For decades the RAM-to-disk memory hierarchy access gap has plagued computer architects. An exciting new storage technology based on microelectromechanical systems (MEMS) is poised to fill a large portion of this performance gap, significantly reduce power consumption, and enable many new classes of applications. This research explores the impact MEMS-based storage will have on computer systems. We examine the performance of several device designs under development. Results from five application studies show these devices reduce application I/O stall times by 3-10X and improve overall application performance by 1.6-8.1X. Further, integrating MEMS-based storage as a disk cache achieves a 3.5X performance improvement over a standalone disk drive. Power consumption simulations show that MEMS devices use up to 10X less power than state-of-the-art low-power disk drives. Many of these improvements stem from the fact that average access times for MEMS-based storage are 10X faster than disks and that MEMS devices are able to rapidly move between active and power-down mode. Combined with the differences in the physical behavior of MEMS-based storage, these characteristics create numerous opportunities for restructuring the storage/memory hierarchy.

Design and Implementation of a Self-Securing Storage Device

Strunk, Goodson, Scheinholtz, Soules & Ganger

CMU SCS Technical Report, CMU-CS-00-129, May 2000.

Self-securing storage prevents intruders from undetectably tampering with or permanently deleting stored data. Self-securing storage devices do this by internally auditing all requests, keeping all versions of all data for a window of time, regardless of the commands received from potentially-compromised host operating systems. Within this window, system administrators have valuable information for intrusion diagnosis and recovery. The S4 implementation combines log structuring with novel metadata journaling and data replication techniques to minimize the performance costs of comprehensive versioning. Experiments show that self-securing storage devices can deliver performance comparable with conventional storage. Further, analyses indicate

... continued on pg. 4

COMINGS & GOINGS

STAFF

Shelby Davis joined the PDL in January as a staff programmer. He is a recent graduate of the CS department.

Nitin Parab travelled from India this spring to join the PDL as a network programmer.

Craig Soules became a staff programmer with the PDL in January. He is continuing his studies at the graduate level with CS in the fall.

Jennifer Landefeld felt the call of PANASAS in January this year and is finding entertainment there as the Operations Manager.

David Rochberg left the PDL in March to work for a start-up here in Pittsburgh.

GRAD STUDENTS

Khalil Amiri will be defending his Ph.D dissertation and graduating this summer from ECE.

Jeff Butler left the PDL at the end of December to work as a programmer at PANASAS.

Fay Chang will also be defending her Ph.D. work and graduating from CS this summer.

Charles Hardin left the PDL at the end of December to join 2Wire, a DSL and home networking development company.

Ed Hogan also left at the end of last year to join PANASAS as a programmer.

Chris Sabol left ECE and the PDL to join Charles at 2Wire this March.

UNDERGRADUATES

Paul Cassella, a senior in CS, joined us as an undergrad programmer at the beginning of January.

Matt Monroe, a senior in CS, left his position as an undergraduate programmer with the PDL this spring to focus on his studies.

... continued from pg. 4

that several weeks worth of all versions can reasonably be kept on state-of-the-art disks, especially when differencing and compression technologies are employed.

Active Disk Architecture for Databases

Riedel, Faloutsos & Nagle

CMU SCS Technical Report, CMU-CS-00-139, May 2000.

Today's commodity disk drives, the basic unit of storage for computer systems large and small, are actually small computers, with a processor, memory and a network connection, in addition to the spinning magnetic material that stores the data. Large collections of data are becoming larger, and people are beginning to analyze, rather than simply store-and-forget, these masses of data. At the same time, advances in I/O performance have lagged the rapid development of commodity processor and memory technology. This paper describes the use of Active Disks to take advantage of the processing power on individual disk drives to run a carefully chosen portion of a relational database system. Moving a portion of the database processing to execute directly at the disk drives improves performance by: 1) dramatically reducing data traffic; and 2) exploiting the parallelism in large storage systems. It provides a new point of leverage to overcome the I/O bottleneck. This paper discusses how to map all the basic database operations - select, project, and join - onto an Active Disk system. The changes required are small and the performance gains are dramatic. A prototype based on the Postgres database system demonstrates a factor of 2x performance improvement on a small system using a portion of the TPC-D decision support benchmark, with the promise of larger improvements in more realistically-sized systems.

Secure Continuous Biometric-Enhanced Authentication

Klosterman & Ganger

CMU SCS Technical Report, CMU-CS-00-134, May 2000.

Biometrics have the potential to solidify person-authentication by examining "unforgeable" features of individuals. This paper explores issues involved with effective integration of biometric-enhanced authentication into computer systems and design options for addressing them. Because biometrics are not secrets, systems must not use them like passwords; otherwise, biometric-based authentication will reduce security rather than increase it. A novel biometric-enhanced authentication system, based on a trusted camera that continuously uses face recognition to verify identity, is described and evaluated in the context of Linux. With cryptographically-signed messages and continuous authentication, the difficulty of bypassing desktop authentication can be significantly increased.

Filling the Memory Access Gap: A Case for On-Chip Magnetic Storage

Schlosser, Griffin, Nagle & Ganger

CMU SCS Technical Report, CMU-CS-99-174, December 1999.

For decades, the memory hierarchy access gap has plagued computer architects with the RAM/disk gap widening to about 6 orders of magnitude in 1999. However, an exciting new storage technology based on MicroElectroMechanical Systems (MEMS) is poised to fill a large portion of this performance gap, delivering significant performance improvements and enabling many new types of applications. This research explores the impact MEMS-based storage will have on computer systems. Working closely with researchers building MEMS-based storage devices, we examine the performance impact of several design points. Results from five different applications show that MEMS-based storage can reduce application I/O stall times by 80-99%, with overall performance improvements ranging from 1.1X to 20X for these applications. Most of these improvements result from the fact that average access times for MEMS-based storage are 5 times faster than disks (e.g., 1-3ms). Others result from fundamental differences in the physical behavior of MEMS-based storage. Combined, these characteristics create numerous opportunities for restructuring the storage/memory hierarchy.

Automated Disk Drive Characterization

Schindler & Ganger

CMU SCS Technical Report, CMU-CS-99-176, December 1999.

DIXtrac is a program that automatically characterizes the performance of modern disk drives. This report describes and validates DIXtrac's algorithms, which extract accurate values for over 100 performance-critical parameters in 2 to 6 minutes without human intervention or special hardware support. The extracted data include detailed layout and geometry information, mechanical timings, cache management policies, and command processing overheads. DIXtrac is validated by configuring a detailed disk simulator with its extracted parameters; in most cases, the resulting accuracies match those of the most accurate disk simulators reported in the literature. DIXtrac has been successfully used on over 20 disk drives, including eight different models from four different manufacturers.