PDL Packet Spring Update

NEWSLETTER ON THE PARALLEL DATA LABORATORY • SPRING 2003

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

CONSORTIUM MEMBERS

EMC Corporation
Hewlett-Packard Laboratories
Hitachi, Ltd.
IBM
Intel Corporation
Microsoft Research
Network Appliance
Panasas, Inc.
Oracle Corporation
Seagate Technology
Sun Microsystems
Veritas Software Corporation

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/

RECENT PUBLICATIONS: ABSTRACTS

Lachesis: Robust Database Storage Management Based on Device-specific Performance Characteristics

Schindler, Ailamaki & Ganger

Carnegie Mellon University School of Computer Science Technical Report CMU-CS-03-124, March 2003.

Database systems work hard to tune I/O performance, but they do not always achieve the full performance potential of modern disk drives. Their abstracted view of storage components hides useful device-specific characteristics, such as disk track boundaries and advanced built-in firmware algorithms. This paper presents a new storage manager architecture, called Lachesis, that exploits a few observable device-specific characteristics to achieve more robust performance. Notably, it enables efficiency nearly equivalent to sequential streaming even in the presence of competing I/O traffic. With automatic adaptation to device characteristics, Lachesis simplifies manual configuration and restores optimizer assumptions about the relative costs of different access patterns expressed in query plans. Based on experiments with both IBM DB2 and an implementation inside the Shore storage manager, Lachesis improves performance of TPC-H queries on average by 10% when running on dedicated hardware. More importantly, it speeds up TPC-H by up to 3x when running concurrently with an OLTP workload, which is simultaneously improved by 7%.

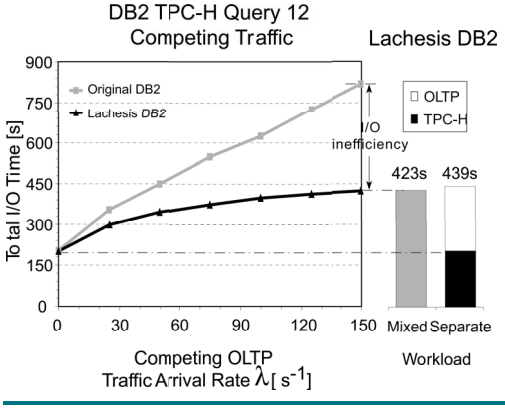
A Case for Staged Database Systems

Harizopoulos & Ailamaki

Proceedings of the 2003 CIDR Conference, Asilomar, CA, January 5-8, 2003.

Traditional database system architectures face a rapidly evolving operating environment, where millions of users store and access terabytes of data. In order to cope with increasing demands for performance, high-end DBMS employ parallel processing techniques coupled with a plethora of sophisticated features. However, the widely adopted, work-centric, thread-parallel execution model entails several shortcomings that limit server performance when executing workloads with changing requirements. Moreover, the monolithic

... continued on pg. 2



TPC-H query 12 execution on DB2. The graph shows the amount of time spent in I/O operations as a function of increasing competing OLTP workload, simulated by random 8 KB I/Os with arrival rate λ. I/O inefficiency in the original DB2 case is due to extraneous rotational delays and disk head switches when running the compound workload. The two bars illustrate the robustness of Lachesis; at each point, both the TPC-H and OLTP traffic achieve their best case efficiency. The *Mixed* workload bar in the right graph corresponds to the λ=150 Lachesis-DB2 datapoint of the left graph. The *Separate* bar adds the total I/O time for the TPC-H and the OLTP-like workloads run separately.

PDL NEWS

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

February 2002

Chenxi Wang Awarded Research Funding from GM

Chenxi Wang has been awarded funding in association with the National Institute of Standards and Technology (NIST) from GM through the General Motors Collaborative Laboratory at Carnegie Mellon to research secure dynamic networks. This contract is a result of GM's recent donation of $8 million over the next five years to Carnegie Mellon University, to continue research on so-called intelligent highways and smart car development. Smart cars process information – such as driver needs and preferences, traffic, road and weather conditions and other information – and make adjustments to avoid travel delays and promote safety.

January 2002

Chris Long and Greg Ganger Receive Funding from C3S

Chris Long and Greg Ganger have been awarded seed funding from the Center for Computer Security (C3S) at Carnegie Mellon for their project "Access Control for the Masses." The project will fall within a new PDL research area dealing with Better User Interfaces.

RECENT PUBLICATIONS

… continued from pg. 1

approach in DBMS software has led to complex and difficult to extend designs. We introduces a staged design for high-performance, evolvable DBMS that are easy to tune and maintain. We propose to break the database system into modules and encapsulate them into self-contained stages connected to each other through queues. The staged, data-centric design remedies the weaknesses of modern DBMS by providing solutions at both a hardware and a software engineering level.

Verifiable Secret Redistribution for Archive Systems

Wong, Wang & Wing

Proceedings of the First International IEEE Security in Storage Workshop, December 2002.

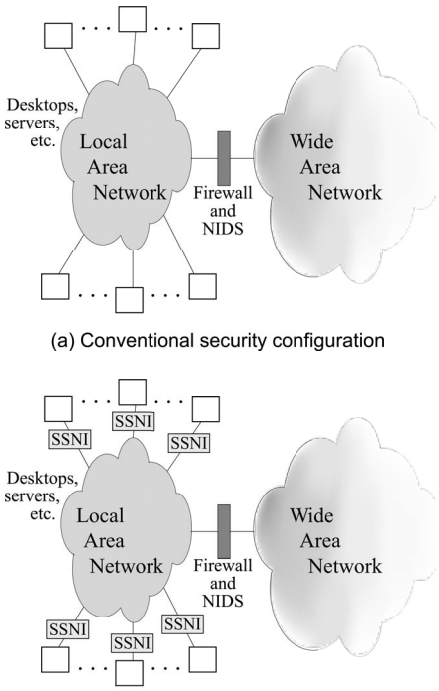
Verifiable secret redistribution is a new protocol for threshold sharing schemes that forms a key component of a proposed archival storage system. Our protocol supports redistribution from (m,n) to (m0,n0) threshold sharing schemes without requiring reconstruction of the original data. The design is motivated by archive systems for which the added security of threshold sharing of data must be accompanied by the flexibility of dynamic shareholder changes. Our protocol enables the dynamic addition or removal of shareholders, and also guards against mobile adversaries. We observe that existing protocols either cannot be extended readily to allow redistribution between different access structures, or have vulnerabilities that allow faulty old shareholders to distribute invalid shares to new shareholders. Our primary contribution is that in our protocol, new shareholders can verify the validity of their shares after redistribution between different access structures.

Finding and Containing Enemies Within the Walls with Self-securing Network Interfaces

Ganger, Economou & Bielski

Carnegie Mellon University School of Computer Science Technical Report CMU-CS-03-109. January 2003.

Self-securing network interfaces (NIs) examine the packets that they move between network links and host software, looking for and potentially blocking malicious network activity. This paper describes how self-securing network interfaces can help administrators to identify and contain compromised machines within their intranet. By shadowing host state, self-securing NIs can better identify suspicious traffic originating from that host, including many explicitly designed to defeat network intrusion detection systems. With normalization and detection-triggered throttling, self-securing



Self-securing network interfaces. (a) shows the common network security configuration, wherein a firewall and a NIDS protect LAN systems from some WAN attacks. (b) shows the addition of self-securing NIs, one for each LAN system.

… continued on pg. 3

… continued from pg. 2

NIs can reduce the ability of compromised hosts to launch attacks on other systems inside (or outside) the intranet. We describe a prototype self-securing NI and example scanners for detecting such things as TTL abuse, fragmentation abuse, "SYN bomb" attacks, and random-propagation worms like Code-Red.

Data Page Layouts for Relational Databases on Deep Memory Hierarchies

Ailamaki, DeWitt & Hill

The VLDB Journal 11(3), 2002.

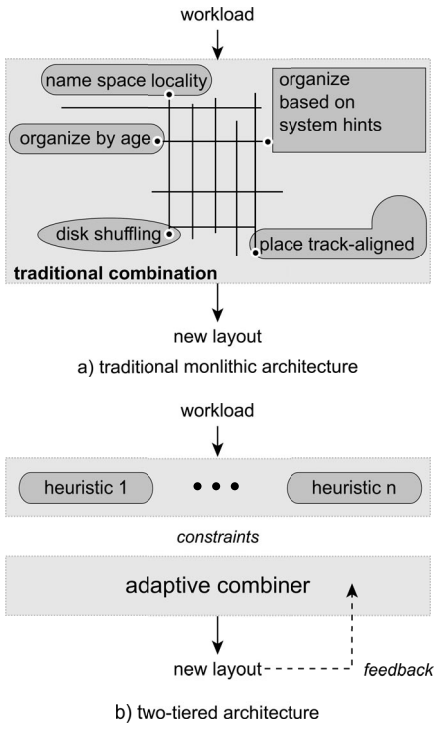
Relational database systems have traditionally optimized for I/O performance and organized records sequentially on disk pages using the N-ary Storage Model (NSM) (a.k.a., slotted pages). Recent research, however, indicates that cache utilization and performance is becoming increasingly important on modern platforms. In this paper, we first demonstrate that in-page data placement is the key to high cache performance and that NSM exhibits low cache utilization on modern platforms. Next, we propose a new data organization model called PAX (Partition Attributes Across), that significantly improves cache performance by grouping together all values of each attribute within each page. Because PAX only affects layout inside the pages, it incurs no storage penalty and does not affect I/O behavior. According to our experimental results (which were obtained without using any indices on the participating relations), when compared to NSM (a) PAX exhibits superior cache and memory bandwidth utilization, saving at least 75% of NSM's stall time due to data cache accesses, (b) range selection queries and updates on memory-resident relations execute 17-25% faster, and (c) TPC-H queries involving I/O execute 11-48% faster. Finally, we show that PAX performs well across different memory system designs.

A Two-tiered Software Architecture for Automated Tuning of Disk Layouts

Salmon, Thereska, Soules & Ganger

Carnegie Mellon School of Computer Science Technical Report CMU-CS-03-130, April 2003.

Many heuristics have been developed for adapting on-disk data layouts to expected and observed workload characteristics. This paper describes a two-tiered software architecture for cleanly and extensibly combining such heuristics. In this architecture, each heuristic is implemented independently and an adaptive combiner merges their suggestions based on how well they work in the given environment. The result is a simpler and more robust system for automated tuning of disk layouts, and a useful blueprint for other complex tuning problems such as cache management, scheduling, data migration, and so forth.



Two-tiered vs. traditional architecture for adaptive layout software. The traditional architecture combines different heuristics in an ad-hoc fashion, usually using a complicated mesh of if-then-else logic. The two-tiered architecture separates the heuristics from the combiner and uses feedback to refine its decisions and utilze the best parts of each heuristic.

A Human Organization Analogy for Self-* Systems

Strunk & Ganger

Carnegie Mellon School of Computer Science Technical Report CMU-CS-03-129, April 2003.

The structure and operation of human organizations, such as corporations, offer useful insights to designers of self-* systems (a.k.a. self-managing or autonomic). This paper explores the analogy, and describes the design of a self-* storage system that borrows from it.

Why Can't I Find My Files? New Methods for Automating Attribute Assignment

Soules & Ganger

Carnegie Mellon University School of Computer Science Technical Report CMU-CS-03-116. February 2003.

This paper analyzes various algorithms for scheduling low priority disk drive tasks. The derived closed form solution is applicable to a class of greedy algorithms that includes a variety of background disk scanning applications. By paying close attention to many characteristics of modern disk drives, the analytical solutions achieve very high accuracy — the difference between the predicted response times and the measurements on two different disks is only 3% for all but one examined workload. This paper also proves a theorem which shows that background tasks implemented by greedy algorithms can be accomplished

… continued on pg. 4

… continued from pg. 3

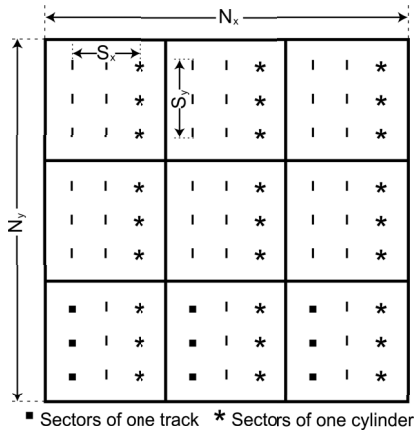
with very little seek penalty. Using greedy algorithm gives a 10% shorter response time for the foreground application requests and up to a 20% decrease in total background task run time compared to results from previously published techniques.

Exposing and Exploiting Internal Parallelism in MEMS-based Storage

Schlosser, Schindler & Ganger

Carnegie Mellon School of Computer Science Technical Report CMU-CS-03-125, March 2003.

MEMS-based storage has interesting access parallelism features. Specifically, subsets of a MEMStore's thousands of tips can be used in parallel, and the particular subset can be dynamically chosen. This paper describes how such access parallelism can be exposed to system software, with minimal changes to system interfaces, and utilized cleanly for two classes of applications. First, background tasks can utilize unused parallelism to access media locations with no impact on foreground activity. Second, two-dimensional data structures, such as dense matrices and relational database tables, can be accessed in both row order and column order with maximum efficiency. With proper table layout, unwanted portions of a table can be skipped while scanning at full speed. Using simulation, we explore performance features of using this device parallelism for an example application from each class.



This figure illustrates the organization of LBNs into tracks and cylinders, and the geometric parameters of the MEMStore. Cylinders are the groups of all LBNs which are at the same offset in the X dimension. Here, all of the LBNs of a sample cylinder are marked as stars. Because the number of LBNs that can be accessed at once is limited by the power budget of the device, a cylinder is accessed sequentially in tracks. The LBNs of a sample track are marked as squares in this picture. Three tracks comprise a single cylinder, since it takes three passes to access an entire cylinder. The parameters Nx and Ny are the number of squares in the X and Y directions, and Sx and Sy are the number of LBNs in a single square in each direction.

Efficient Consistency for Erasure-coded Data Via Versioning Servers

Goodson, Wylie, Ganger & Reiter

Carnegie Mellon University School of Computer Science Technical Report CMU-CS-03-127, April 2003.

This paper describes the design, implementation and performance of a family of protocols for survivable, decentralized data storage. These protocols exploit storage-node versioning to efficiently achieve strong consistency semantics. These protocols allow erasure-codes to be used to achieve network and storage efficiency (and optionally data confidentiality in the face of server compromise). The protocol family is general in that its parameters accommodate a wide range of fault and timing assumptions, up to asynchrony and Byzantine faults of both storage-nodes and clients, with no changes to server implementation or client-server interface. Measurements of a prototype storage system using these protocols show that the protocol performs well under various system model assumptions, numbers of failures tolerated, and degrees of reader-writer concurrency.

The DiskSim Simulation Environment Version 3.0 Reference Manual

Bucy, Ganger & Contributors

Carnegie Mellon University School of Computer Science Technical Report CMU-CS-03-102, January 2003.

DiskSim is an efficient, accurate and highly-configurable disk system simulator developed to support research into various aspects of storage subsystem architecture. It includes modules that simulate disks, intermediate controllers, buses, device drivers, request schedulers, disk block caches, and disk array data organizations. In particular, the disk drive module simulates modern disk drives in great detail and has been carefully validated against several production disks (with accuracy that exceeds any previously reported simulator).

This manual describes how to configure and use DiskSim, which has been made publicly available with the hope of advancing the state-of-the-art in disk system performance evaluation in the research community. The manual also briefly describes DiskSim's internal structure and various validation results.

Graduate students discussing research with Microsoft's Bruce Worthington.