PDL Packet Spring Update

NEWSLETTER ON PDL ACTIVITIES AND EVENTS • SPRING 2007
http://www.pdl.cmu.edu/

PDL CONSORTIUM MEMBERS

American Power Conversion
Cisco Systems
EMC
Google
Hewlett-Packard Labs
Hitachi
IBM
Intel
LSI
Network Appliance
Oracle
Panasas
Seagate Technology
Symantec

CONTENTS

THE PDL PACKET

EDITOR
Joan Digney
CONTACTS
Greg Ganger
PDL Director
Bill Courtright
PDL Executive 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

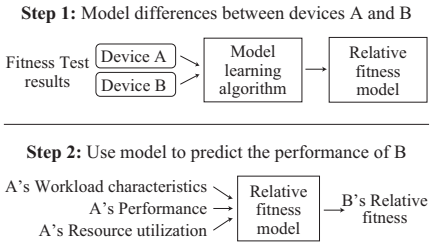
Modeling the Relative Fitness of Storage

Mesnier, Wachs, Sambasivan, Zheng & Ganger

SIGMETRICS'07, June 12–16, 2007, San Diego, California, USA.

Relative fitness is a new black-box approach to modeling the performance of storage devices. In contrast with an absolute model that predicts the performance of a workload on a given storage device, a relative fitness model predicts performance *differences* between a pair of devices. There are two primary advantages to this approach. First, because a relative fitness model is constructed for a device pair, the application-device feedback of a closed workload can be captured (e.g., how the I/O arrival rate changes as the workload moves from device A to device B). Second, a relative fitness model allows performance and resource utilization to be used in place of workload characteristics. This is beneficial when workload characteristics are difficult to obtain or concisely express (e.g., rather than describe the spatio-temporal characteristics of a workload, one could use the observed cache behaviour of device A to help predict the performance of B).

This paper describes the steps necessary to build a relative fitness model, with an approach that is general enough to be used with any black-box modeling technique. We compare relative fitness models and absolute models across a variety of workloads and storage devices. On average, relative fitness models predict bandwidth and throughput within 10–20% and can reduce prediction error by as much as a factor of two when compared to absolute models.



Relative fitness models predict changes in performance between two devices; an I/O "fitness test" is used to build a model of the performance differences. For new workloads, one inputs into the model the workload characteristics, performance, and resource utilization of a workload on device A to predict the relative fitness of device B.

Consistency-preserving Caching of Dynamic Database Content

Tolia & Satyanarayanan

International World Wide Web Conference (WWW 2007), May 8-12, 2007, Banff, Alberta, Canada.

With the growing use of dynamic web content generated from relational databases, traditional caching solutions for throughput and latency improvements are ineffective. We describe a middleware layer called Ganesh that reduces the volume of data transmitted without semantic interpretation of queries or results. It achieves this reduction through the use of cryptographic hashing to detect similarities with previous results. These benefits do not require any compromise of the

continued on page 2

PDL NEWS & AWARDS

April 2007

Elie Krevat Awarded NDSEG Fellowship

Congratulations to Elie Krevat, who has been selected to receive a 2007 National Defense Science and Engineering Graduate (NDSEG) Fellowship. The NDSEG Fellowship is sponsored and funded by the Department of Defense (DoD). NDSEG selections were made from a pool of more than 3,400 applications by the Air Force Research Laboratory/Air Force Office of Scientific Research (AFRL/AFOSR), the Office of Naval Research (ONR), the Army Research Office (ARO), and the DoD High Performance Computing Modernization Program Office (HPCMP). The NDSEG Fellowship covers tuition and required fees for three years at any accredited U.S. college or university that offers advanced degrees in science and engineering. In addition, the NDSEG Fellowship will provide a yearly stipend.

April 2007

Photo Exhibit by PDL Student

Eno Thereska, who during graduate school led a parallel life as a photographer, has a photo exhibition running from April–June. The exhibition, titled "Species," is showing at the Pittsburgh Filmmakers gallery.

March 2007

PDL Researchers Awarded Best Paper at SIGMETRICS 2007

The program chairs of SIGMETRICS 2007, which will be held from June 12-16 in San Diego, CA, have announced that the Best Paper Award will be given to a team of researchers from the Parallel Data Lab (PDL) for their work, "Modeling the Relative Fitness of Storage." The authors are graduate students Michael Mesnier (ECE), Matthew Wachs (CS), Raja Sambasivan (ECE), CS postdoctoral research fellow Alice Zheng, and their faculty advisor, Greg Ganger, PDL director and Professor of ECE and CS.

The paper was chosen from among the 29 accepted (and many others submit-

continued on page 6

RECENT PUBLICATIONS

continued from page 1

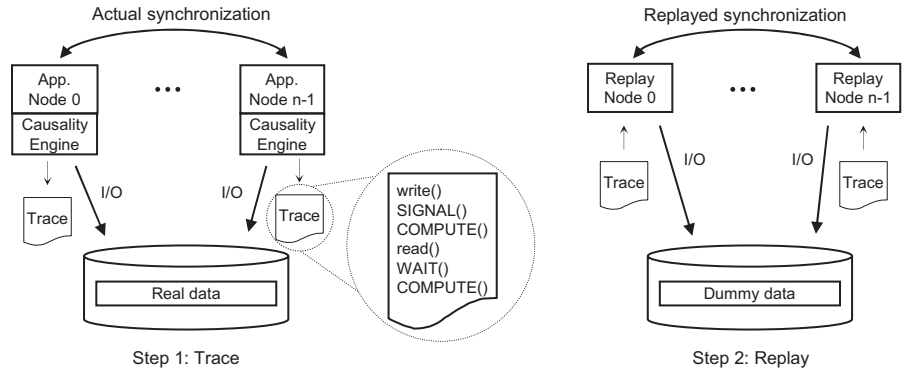
strict consistency semantics provided by the back-end database. Further, Ganesh does not require modifications to applications, web servers, or database servers, and works with closed-source applications and databases. Using two benchmarks representative of dynamic web sites, measurements of our prototype show that it can increase end-to-end throughput by as much as twofold for non-data intensive applications and by as much as tenfold for data intensive ones.

//TRACE: Parallel Trace Replay with Approximate Causal Events

Mesnier, Wachs, Sambasivan, Lopez, Hendricks & Ganger

Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07), February 13–16, 2007, San Jose, CA.

//TRACE (pronounced parallel trace) is a new approach for extracting and replaying traces of parallel applications to recreate their I/O behavior.



//Trace High-level architecture. While an application is running (left half of figure) the nodes are traced by the causality engine (a dynamically linked library) and selectively throttled to expose their data dependencies. Computation times are also estimated. This information is then used to annotate the I/O traces with SIGNAL(), WAIT() and COMPUTE() calls that can be easily replayed in a distributed replayer (right half of figure). During replay, dummy data files are use in place of the real data files.

Its tracing engine automatically discovers inter-node data dependencies and inter-I/O compute times for each node (process) in an application. This information is reflected in per-node annotated I/O traces. Such annotation allows a parallel replayer to closely mimic the behavior of a traced application across a variety of storage systems. When compared to other replay mechanisms, //TRACE offers significant gains in replay accuracy. Overall, the average replay error for the parallel applications evaluated in this paper is below 6%.

continued on page 3

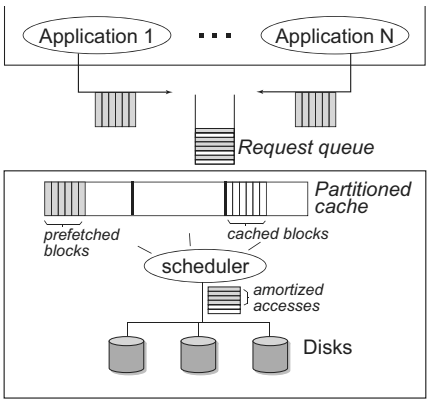
continued from page 2

Argon: Performance Insulation for Shared Storage Servers

Wachs, Abd-El-Malek, Thereska, & Ganger.

Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07), February 13–16, 2007, San Jose, CA.

Services that share a storage system should realize the same efficiency, within their share of time, as when they have the system to themselves. The Argon storage server explicitly manages its resources to bound the inefficiency arising from inter-service disk and cache interference in traditional systems. The goal is to provide each service with at least a configured fraction (e.g., 0.9) of the throughput it achieves when it has the storage server to itself, within its share of the server—a service allocated 1 nth of a server should get nearly 1 nth (or more) of the throughput it would get alone. Argon uses automatically configured prefetch/write-back sizes to insulate streaming efficiency from disk seeks introduced by competing workloads. It uses explicit disk time quanta to do the same for non-streaming workloads with internal locality. It partitions the cache among services, based on their observed access patterns, to insulate the hit rate each achieves from the access patterns of others. Experiments show that, combined, these mechanisms and Argon's automatic configuration of each achieve the insulation goal.



Argon's high-level architecture. Argon makes use of cache partitioning, request amortization, and quanta-based disk time scheduling.

Observer: Keeping System Models from Becoming Obsolete

Thereska, Narayanan, Ailamaki, & Ganger.

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-07-101, January 2007.

To be effective for automation, in practice, system models used for performance prediction and behavior checking must be robust. They must be able to cope with component upgrades, misconfigurations, and workload-system interactions that were not anticipated. This paper promotes making models self-evolving, such that they continuously evaluate their accuracy and adjust their predictions accordingly. Such self-evaluation also enables confidence values to be provided with predictions, including identification of situations where no trustworthy prediction can be produced. With a combination of expectation-based and observation-based techniques, we believe that such self-evolving models can be achieved and used as a robust foundation for tuning, problem diagnosis, capacity planning, and administration tasks.

Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?

Schroeder & Gibson

Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST '07), February 13–16, 2007, San Jose, CA.

Component failure in large-scale IT installations such as cluster supercomputers or internet service providers is becoming an ever larger problem as the number of processors, memory chips and disks in a single cluster approaches a million. In this paper, we present and analyze field-gathered disk replacement data from five systems in production use at three organizations, two supercomputing sites and one internet service provider. About 70,000 disks are covered by this data, some for an entire lifetime of 5 years. All disks were high-performance enterprise disks (SCSI or FC), whose datasheet MTTF of 1,200,000 hours suggest a nominal annual failure rate of at most 0.75%.

Raja takes a short break from the proceedings of the 2006 PDL Workshop & Retreat.

We find that in the field, annual disk replacement rates exceed 1%, with 2-4% common and up to 12% observed on some systems. This suggests that field replacement is a fairly different process than one might predict based on datasheet MTTF, and that it can be quite variable installation to installation.

We also find evidence that failure rate is not constant with age, and that rather than a significant infant mortality effect, we see a significant early onset of wear-out degradation. That is, replacement rates in our data grew constantly with age, an effect often assumed not to set in until after 5 years of use.

In our statistical analysis of the data, we find that time between failure is not well modeled by an exponential distribution, since the empirical distribution exhibits higher levels of variability and decreasing hazard rates. We also find significant levels of correlation between failures, including autocorrelation and long-range dependence.

continued on page 4

continued from page 3

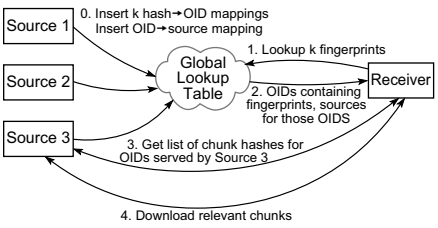
Exploiting Similarity for Multi-Source Downloads using File Handprints

Pucha, Andersen, & Kaminsky

Proceedings of the 4th USENIX NSDI, Cambridge, MA, April 2007.

Many contemporary approaches for speeding up large file transfers attempt to download chunks of a data object from multiple sources. Systems such as BitTorrent quickly locate sources that have an exact copy of the desired object, but they are unable to use sources that serve similar but non-identical objects. Other systems automatically exploit cross-file similarity by identifying sources for each chunk of the object. These systems, however, require a number of lookups proportional to the number of chunks in the object and a mapping for each unique chunk in every identical and similar object to its corresponding sources. Thus, the lookups and mappings in such a system can be quite large, limiting its scalability.

This paper presents a hybrid system that provides the best of both approaches, locating identical and similar sources for data objects using a constant number of lookups and inserting a constant number of mappings per object. We first demonstrate through extensive data analysis that similarity does exist among objects of popular file types, and that making use of it can sometimes substantially improve download times. Next, we describe handprinting, a technique that allows clients to locate similar sources using a constant number of lookups and mappings. Finally, we describe the design, implementation and evaluation of Similarity-Enhanced Transfer (SET), a system that uses this technique to download objects. Our experimental evaluation shows that by using sources of similar objects, SET is able to significantly out-perform an equivalently configured BitTorrent.



SET design overview

John Wilkes, of HP Labs, discusses "Prato: A Virtual-DBMS-Appliance Service Provider" with PDL students at an industry poster session held in conjunction with the 2006 PDL Workshop & Retreat. This poster session was designed to give retreat participants an opportunity to see what sort of projects our Consortium Members are working on.

Fingerpointing Correlated Failures in Replicated Systems

Pertet, Gandhi & Narasimhan

USENIX Workshop on Tackling Computer Systems Problems with Machine Learning Techniques (SysML), Cambridge, MA (April 2007).

Replicated systems are often hosted over underlying group communication protocols that provide totally ordered, reliable delivery of messages. In the face of a performance problem at a single node, these protocols can cause correlated performance degradations at even non-faulty nodes, leading to potential red herrings in failure diagnosis. We propose a fingerpointing approach that combines node-level (local) anomaly detection, followed by system-wide (global) fingerpointing. The local anomaly detection relies on threshold-based analyses of system metrics, while global fingerpointing is based on the hypothesis that the root-cause of the failure is the node with an "odd-man-out" view of the anomalies. We compare the results of applying three classifiers – a heuristic algorithm, an unsupervised learner (k-means clustering), and a supervised learner (k-nearest-neighbor) – to fingerpoint the faulty node.

MultiMap: Preserving Disk Locality for Multidimensional Datasets

Shao, Papadomanolakis, Schlosser, Schindler, Ailamaki & Ganger

IEEE 23rd International Conference on Data Engineering (ICDE 2007) Istanbul, Turkey, April 2007.

MultiMap is an algorithm for mapping multidimensional datasets so as to preserve the data's spatial locality on disks. Without revealing disk-specific details to applications, MultiMap exploits modern disk characteristics to provide full streaming bandwidth for one (primary) dimension and maximally efficient non-sequential access (i.e., minimal seek and no rotational latency) for the other dimensions. This is in contrast to existing approaches, which either severely penalize non-primary dimensions or fail to provide full streaming bandwidth for any dimension. Experimental evaluation of a prototype implementation demonstrates MultiMap's superior performance for range and beam queries. On average, MultiMap reduces total I/O time by over 50% when compared to traditional linearized layouts and by over 30% when compared to space-filling curve approaches such as Z-ordering and Hilbert curves. For scans of the primary dimension, MultiMap and traditional linearized layouts provide almost two orders of magnitude higher throughput than space-filling curve approaches.

continued on page 8

DISSERTATION ABSTRACT:
Efficient Cryptographic Techniques for Securing Storage Systems

Alina Mihaela Oprea

Carnegie Mellon University School of Computer Science Ph.D. Dissertation, January 18, 2007.

Networked storage solutions, such as Network-Attached Storage and Storage Area Networks, are recently emerging storage architectures that provide higher performance and availability than traditional direct-attached disks. In these environments, the networked storage devices are subject to attacks, and, consequently, clients have to play a more proactive role in ensuring the confidentiality and integrity of data. For securing the data stored remotely, we consider an architecture in which clients have access to a small amount of trusted storage, which could either be local to each client or, alternatively, could be provided by a client's organization through a dedicated server.

In this thesis, we propose new approaches for three different mechanisms that are currently employed in implementations of secure storage systems. In designing the new algorithms for securing storage systems, we set three main goals. First, security should be added by clients transparently for the storage servers so that the storage interface does not change; second, the amount of trusted storage used by clients should be reduced; and, third, the performance overhead of the security algorithms should not be prohibitive.

The first contribution of this dissertation is the construction of novel space-efficient integrity algorithms for both block-level storage systems and cryptographic file systems. These constructions are based on the observation that block contents typically written to disks feature low entropy, and as such are efficiently distinguishable from uniformly random blocks. We provide a rigorous analysis of security of the new integrity algorithms and demonstrate that they maintain the same security properties as existing algorithms (e.g., Merkle tree). We implement the new algorithms for integrity checking of files in the EncFS cryptographic file system and measure their performance cost, as well as the amount of storage needed for integrity and the integrity bandwidth (i.e., the amount of information needed to update or check the integrity of a file block) used. We evaluate the block-level integrity algorithms using a disk trace we collected, and the integrity algorithms for file systems using NFS traces collected at Harvard university and a file trace from a standard Linux distribution.

We also construct efficient key management schemes for cryptographic file systems in which the re-encryption of a file following a user revocation is delayed until the next write to that file, a model called lazy revocation. The encryption key evolves at each revocation and we devise an efficient algorithm to recover previous encryption keys with only logarithmic cost in the number of revocations supported. The novel key management scheme is based on a binary tree to derive the keys and improves existing techniques by several orders of magnitude, as shown by our experiments.

Alina Oprea discusses "Integrity Checking in Cryptographic File Systems with Constant Trusted Storage" at the 2006 PDL Workshop and Retreat.

Our final contribution is to analyze theoretically the consistency of encrypted shared file objects used to implement cryptographic file systems. We provide sufficient conditions for the realization of a given level of consistency, when concurrent writes to both the file and encryption key objects are possible. We show that the consistency of both the key distribution and the file access protocol affect the consistency of the encrypted file object that they implement. To demonstrate that our framework simplifies complex proofs for showing the consistency of an encrypted file, we provide a simple implementation of a fork consistent encrypted file and prove its consistency.

THESIS PROPOSAL:
Delayed Instantiation Bulk Operations in a Clustered, Object-based Storage System

Andrew J. Klosterman, ECE

Many storage management tasks contain, at their heart, a step that applies the same operation to many stored data items: a bulk operation. Such bulk operations have evolved over the years in enterprise-class block- and file-based storage systems. A new breed of storage, object-based storage, will benefit from supporting bulk operations that have come to be expected in established storage systems.

This thesis proposes the investigation of ways to support the execution of specific management tasks on flexibly defined sets of objects in a clustered, object-based storage system. The semantics and performance of such tasks are expected to at least meet, and hopefully exceed, those of support-

continued on page 7

continued from page 2

ted) for publication at the conference, which focuses on the measurement and modeling of computer systems.

SIGMETRICS is the Association for Computing Machinery's (ACM) Special Interest Group (SIG) for computer/communication system performance. SIGMETRICS promotes research in performance analysis techniques as well as the advanced and innovative use of known methods and tools.

March 2007

Mike Kasick Awarded NSF Graduate Fellowship

ECE student Mike Kasick has been awarded an NSF Graduate Research Fellowship. CMU had 8 awardees in all, with only one from ECE.

The fellowship provides funding for a maximum of three years that can be used over a five-year period, including a stipend of $30,000 per twelve-month fellowship period. Mike is advised in his research on problem diagnosis by Priya Narasimhan.

February 2007

Two PDL Researchers Awarded Sloan Fellowships

Two Sloan Fellowships in computer science have been awarded to PDL faculty members: Priya Narasimhan, ECE and ISR, and Dawn Song, ECE and CSD.

A Sloan Fellowship is a prestigious award intended to enhance the careers of the very best young faculty members in specified fields of science. Currently a total of 118 fellowships $45,000 fellowships each year are awarded annually in seven fields: chemistry, computational and evolutionary molecular biology, computer science, economics, mathematics, neuroscience, and physics. Only 16 are given in computer science each year so CMU once again shines. Since the establishment of the fellowships in 1955, 32 Sloan Fellows have gone on to win Nobel Prizes.

February 2007

PDL Researchers Win Best Paper at FAST 2007!

Bianca Schroeder and Garth Gibson have brought home the Best Paper Award from the 5th USENIX Conference on File and Storage Technologies (FAST 2007), which was held in San Jose, CA this year. The award was given for their research on "Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?"

PDL researchers have been well received at past FAST conferences as well, winning best student paper awards in 2002 ("Track-Aligned Extents: Matching Access Patterns to Disk Drive Characteristics", Schindler, et al.) and in 2004 ("A Framework for Building Unobtrusive Disk Maintenance Applications", Thereska et al.). In 2005 the PDL brought home both best paper awards for work on "Ursa Minor: Versatile Cluster-based Storage", Abd-El-Malek et al. and "On Multidimensional Data and Modern Disks", Schlosser et al.

January 2007

Dawn Song Selected for College of Engineering Award

Dawn Song is among the three ECE faculty members who won awards from the College of Engineering this year. Song, Assistant Professor of ECE and Computer Science, received a George Tallman Ladd Research Award, which is granted in recognition of outstanding research, professional accomplishments, and potential.

-- with info from ECE News Online

December 2006

Michael Kasick Honored by Computing Research Association

Michael Kasick, a senior in ECE, was selected as a finalist in the Computing Research Association's (CRA) Outstanding Undergraduate Award competition for 2007. The annual award recognizes undergraduates from North American universities who show outstanding potential in computing research.

Kasick became interested in conducting undergraduate research at Carnegie Mellon after taking Embedded Real-Time Systems, taught by Priya Narasimhan, Assistant Professor of ECE and Institute for Software Research (ISR). He began by volunteering with her in the summer of 2005 on a project investigating the research challenges underlying fingerpointing (also know as root-cause analysis or failure diagnosis) in large-scale distributed systems.

Later, his CIT honors project involved developing algorithms and tools to assist administrators of the real-world Emulab 400+-node cluster (located at the University of Utah) in diagnosing the root cause of failures. His results, "Towards Fingerpointing in the Emulab Dynamic Distributed System," were published last month at the USENIX Workshop on Real Large Distributed Systems (WORLDS), which was held in Seattle, in conjunction with the

continued on page 7

continued from page 6

USENIX Symposium on Operating Systems Design and Implementation (OSDI).

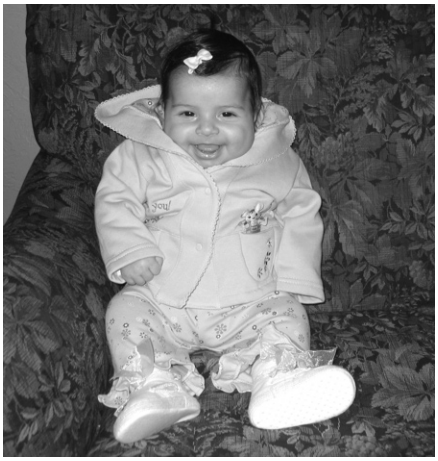
"I have advised many undergraduate and graduate students, and I have to admit that Mike's prolific progress in his research has blown me away," said Narasimhan. "It is very hard for even graduate students to get their research work published at a workshop like WORLDS. Great things await Mike in the future, and I am excited and privileged to be along for the ride."

Outside of the classroom, Kasick serves in leadership roles in Carnegie Mellon's Computer Club and the Carnegie Tech Radio Club (W3VC). He has been admitted into the Ph.D. program in ECE, and plans to continue his fingerpointing research.

Students are nominated for the CRA Outstanding Undergraduate Award by their department. Other honorees from Carnegie Mellon include: winner Stephanie Rosenthal, a senior in Computer Science (CS) and Human-Computer Interaction (HCI); finalist Mihir Kedia, a senior in CS; and honorable mention nominee Brendan Meeder, a senior in CS and Math.

An announcement of the winners will appear in the January 2007 issue of Computing Research News and the awards will be presented at an upcoming computing research conference. This year's award program is sponsored by Microsoft Research.

-- Source: Computing Research Assoc.



December 2006

A New Member of the Courtright Family!

Theresa Anne Courtright arrived at 10:05 am, on Dec. 9th 2006, weighing 7 lbs. 11 oz., and measuring 20 inches in length. Her proud family are Dad and Mom, Bill and Mireille Courtright and big brother William. Congratulations to all!

November 2006

Carnegie Mellon Researchers Win HPC Analytics Challenge at ACM/IEEE SC2006

Computer science (CS) graduate student Tiankai Tu and David O'Hallaron, Associate Professor of CS and Electrical and Computer Engineering (ECE), led a team of researchers to win the High Performance Computing (HPC) Analytics Challenge at ACM/IEEE Supercomputing 2006 in Tampa, FL. ECE graduate student Julio Lopez was also a member of the winning group.

The entry is titled Remote Runtime Steering of Integrated Terascale Simulation and Visualization. The team developed a novel analytic capability that enables scientists and engineers to obtain insights from on-going large-scale parallel unstructured finite element mesh simulations. During the Analytics Challenge session, the team showed a live demo: steering, in real-time, the visualization of a 2050-processor earthquake ground motion simulation running on the Cray XT3 supercomputer in Pittsburgh, PA, via a wireless Internet connection, from a laptop computer in the conference room in Tampa, FL.

The Carnegie Mellon team members were Tiankai Tu (team lead), Jacobo Bielak, Julio Lopez, David O'Hallaron, Leonardo Ramirez-Guzman, and Ricardo Taborda-Rios. The other team members were: Hongfeng Yu (technical lead) and Kwan-Liu Ma of the University of California, Davis; Omar Ghattas of the University of Texas at Austin; and Nathan Stone and John Urbanic of the Pittsburgh Supercomputing Center.

-- Source: Byron Spice, Carnegie Mellon Computer Science News

continued from page 5

ing operations on current block- and file-based storage systems. This is to be done while coping with the challenges presented by the distributed nature of storage in clustered object-based storage systems.

Through the delayed instantiation of bulk operations on objects, performance will be maintained and operation semantics upheld. The use of copy-on-write and lazy evaluation techniques, along with capability-based access control, enables the use of delayed instantiation.

By investigating the effects of delayed instantiation on different workloads and client-access scenarios, the associated costs can be measured and compared in a prototype clustered, object-based storage system. Furthermore, advantageous structuring of higher-level storage systems built atop the prototype will be demonstrated and characterized.

continued from page 4

Database Servers on Chip Multiprocessors: Limitations and Opportunities

Hardavellas, Pandis, Johnson, Mancheril, Ailamaki & Falsafi

Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, January 2007.

Prior research shows that database system performance is dominated by off-chip data stalls, resulting in a concerted effort to bring data into on-chip caches. At the same time, high levels of integration have enabled the advent of chip multiprocessors and increasingly large (and slow) on-chip caches. These two trends pose the imminent technical and research challenge of adapting high-performance data management software to a shifting hardware landscape. In this paper we characterize the performance of a commercial database server running on emerging chip multiprocessor technologies. We find that the major bottleneck of current software is data cache stalls, with L2 hit stalls rising from oblivion to become the dominant execution time component in some cases. We analyze the source of this shift and derive a list of features for future database designs to attain maximum performance.

Scheduling Threads for Constructive Cache Sharing on CMPs

Chen, Gibbons, Kozuch, Liaskovitis, Ailamaki, Blelloch, Falsafi, Fix, Hardavellas, Mowry & Wilkerson

19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA' 07), San Diego, CA, June 2007.

In chip multiprocessors (CMPs), limiting the number of off-chip cache misses is crucial for good performance. Many multithreaded programs provide opportunities for *constructive* cache sharing, in which concurrently scheduled threads share a largely overlapping working set. In this paper, we compare the performance of two state-of-the-art schedulers proposed for ne-grained multithreaded programs: Parallel Depth First (PDF), which is specifically designed for constructive cache sharing, and Work Stealing (WS), which is a more traditional design. Our experimental results indicate that PDF scheduling yields a 1.3–1.6X performance improvement relative to WS for several fine-grain parallel benchmarks on projected future CMP configurations; we also report several issues that may limit the advantage of PDF in certain applications. These results also indicate that PDF more effectively utilizes off-chip bandwidth, making it possible to trade-off on-chip cache for a larger number of cores. Moreover, we find that task granularity plays a key role in cache performance. Therefore, we present an automatic approach for selecting effective grain sizes, based on a new working set profiling algorithm that is an order of magnitude faster than previous approaches. This is the first paper demonstrating the effectiveness of PDF on real benchmarks, providing a direct comparison between PDF and WS, revealing the limiting factors for PDF in practice, and presenting an approach for overcoming these factors.

Using Provenance to Aid in Personal File Search

Shah, Soules, Ganger & Noble

Usenix Annual Technical Conference, Santa Clara, CA, June 17–22, 2007.

As the scope of personal data grows, it becomes increasingly difficult to find what we need when we need it. Desktop search tools provide a potential answer, but most existing tools are incomplete solutions: they index content, but fail to capture dynamic relationships from the user's context. One emerging solution to this is context-enhanced search, a technique that reorders and extends the results of content-only search using contextual information. Within this framework, we propose using strict *causality*, rather than temporal locality, the current state of the art, to direct contextual searches. Causality more accurately identifies data flow between files, reducing the false-positives created by context-switching and background noise. Further, unlike previous work, we conduct an online user study with a fully-functioning implementation to evaluate *user-perceived* search quality directly. Search results generated by our causality mechanism are rated a statistically-significant 17% higher, on average over all queries, than by using content-only search or context-enhanced search with temporal locality.

Improving Mobile Database Access Over Wide-Area Networks Without Degrading Consistency

Tolia, Satyanarayanan & Wolbach

MobiSys '07, June 11–13, 2007, San Juan, Puerto Rico, USA.

We report on the design, implementation, and evaluation of a system called Cedar that enables mobile database access with good performance over low-bandwidth networks. This is accomplished without degrading consistency. Cedar exploits the disk storage and processing power of a mobile client to compensate for weak connectivity. Its central organizing principle is that even a stale client replica can be used to reduce data transmission volume from a database server. The reduction is achieved by using content addressable storage to discover and elide commonality between client and server results. This organizing principle allows Cedar to use an optimistic approach to solving the difficult problem of database replica control. For laptop-class clients, our experiments show that Cedar improves the throughput of read-write workloads by 39% to as much as 224% while reducing response time by 28% to as much as 79%.