# PDL Packet

NEWSLETTER ON PDL ACTIVITIES AND EVENTS • FALL 2005
http://www.pdl.cmu.edu/

PARALLEL DATA LABORATORY
CARNEGIE MELLON UNIVERSITY

AN INFORMAL PUBLICATION FROM ACADEMIA'S PREMIERE STORAGE SYSTEMS RESEARCH CENTER DEVOTED TO ADVANCING THE STATE OF THE ART IN STORAGE AND INFORMATION INFRASTRUCTURES.

## CONTENTS

## PDL CONSORTIUM MEMBERS

American Power Corporation
EMC Corporation
EqualLogic, Inc.
Hewlett-Packard Labs
Hitachi, Ltd.
IBM Corporation
Intel Corporation
Microsoft Corporation
Network Appliance
Oracle Corporation
Panasas, Inc.
Seagate Technology
Sun Microsystems

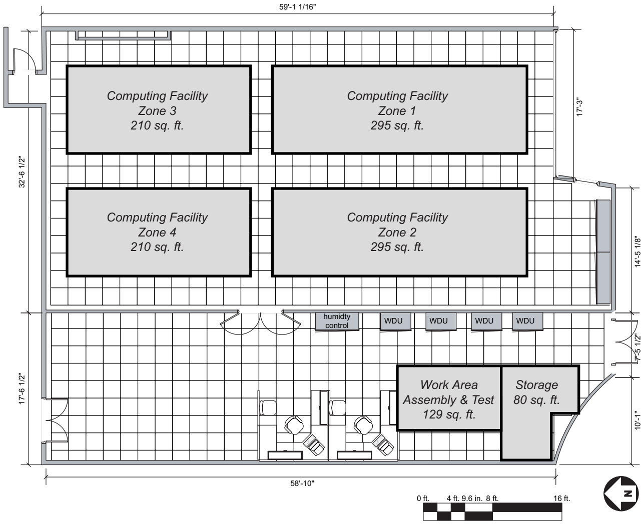
# Data Center Observatory

*Greg Ganger & Bill Courtright*

The Data Center Observatory (DCO) was conceived to guide and enable experimental research into data center automation and efficiency. The DCO will be a fully-instrumented operational data center at Carnegie Mellon, simultaneously used as a utility and as a research vehicle. Its computation and storage resources will be a utility for CMU research activities, from data mining to CAD/architecture to visualization, and real networked services. At the same time, PDL researchers will measure all operational aspects, including power/thermal, human administrator time, resource utilization patterns, failures and their consequences, and so on. Further, technologies and tools developed to mitigate operational costs will be deployed in this real environment, and measurements will quantify their efficacy and guide refinement in an interative research cycle.

## Why a Data Center Observatory is Needed

There is broad agreement that human administration and other operational costs are major issues for data centers—perhaps the biggest challenge facing modern computing and certainly the dominant factor in total cost of ownership. But, there is little specific information about where administration time and effort goes and the breakdown of operational costs. As a consequence, it is difficult to focus researchers and engineers on the real problems—it is difficult to solve a problem that cannot be specified. Further, for many known administration challenges, such as diagnosis of performance problems or application failures, it is difficult to evaluate the effectiveness of a proposed solution. For better or worse, "administratability" does not appear to lend itself to controlled lab experiments.



Floor plan of the Data Center Obersvatory.

The data center observatory was conceived as an aggressive, but necessary, case study for understanding operational costs and evaluating solutions. Only by measuring and characterizing operational costs in real environments can we focus research attention on the most critical unsolved aspects. Explicitly instrumented to act as an observatory, the DCO will be one such real environment. We hope to deploy proven instrumentation technologies,

*continued on page 10*

## FROM THE DIRECTOR'S CHAIR

### Greg Ganger

Hello from fabulous Pittsburgh!

2005 has been an exciting year for the Parallel Data Lab, with the return of two key players and investment in infrastructure build-up that will enable the next several years of PDL research. Along the way, a new faculty member and several new students have joined PDL, several students have graduated and taken jobs with PDL Consortium companies, and many papers have been published.

For me, the most exciting thing has been the returns of Garth Gibson and Bill Courtright. Of course, Garth founded the PDL 13 years ago and has been with us in spirit during his extended leave of absence from Carnegie Mellon. He is now back, about half-time with the other half at Panasas, teaching classes and starting into research activities again. His initial foci have been pNFS and cluster management, and all of us look forward to working with him more and more. Bill Courtright has a long PDL history, starting as a Ph.D. student in PDL's first year and returning as Executive Director from 1998-1999. He has again returned as Executive Director, after co-founding Panasas, and his organizational help has been a godsend for me.

The PDL continues to pursue a broad array of storage systems research, and this past year brought much progress on the exciting new projects launched last year. Let me highlight a few things.

Of course, first up is the our primary umbrella project, Self-* Storage, which explores the design and implementation of self-organizing, self-configuring, self-tuning, self-healing, self-managing systems of storage bricks. For years, PDL Retreat attendees pushed us to attack "storage management of large installations", and this project is our response. In early project planning stages, after many discussions with PDL Consortium members and real IT staffs, we came to the conclusion that support for manageability must be designed into storage architecture from the beginning, rather than added after the fact. So, we have given ourselves a clean slate. Our initial design (a system dubbed Ursa Major) combines features of several PDL projects (e.g., PASIS and self-securing storage) with heavy doses of instrumentation and agents for processing observations and enacting decisions. As a first step, we have constructed Ursa Minor, which is a versatile cluster storage system that can act as the base for Ursa Major. With the Ursa Minor foundation, research into many components of achieving self-*-ness is being attacked, including performance insulation, performance instrumentation and predictability, metadata scalability and recoverability, multi-metric goal-based tuning, and new approaches to device modeling.

One of the biggest challenges in a project like Self-* Storage is evaluation -- lab experiments simply do not capture the unpredictable occurrences, each instance unique, of real deployments. We have decided that the only way to go is to dive in fully and deploy a real system to test our ideas in practice. This will allow us to see, first-hand, which challenges are and are not being successfully addressed and then iterate accordingly. With generous equipment donations from the PDL Consortium companies, combined with government grants, we hope to put together and maintain hundreds of terabytes of storage used by ourselves and other CMU researchers (e.g., in data mining, CAD, and scientific visualization).

THE PDL PACKET

The Parallel Data Laboratory
School of Computer Science
Department of ECE
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
VOICE 412•268•6716
FAX 412•268•3010

PUBLISHER
Greg Ganger

EDITOR
Joan Digney

The PDL Packet is published once per year to update members of the PDL Consortium. A pdf version resides in the Publications section of the PDL Web pages and may be freely distributed. Contributions are welcome.

COVER ILLUSTRATION

Skibo Castle and the lands that comprise its estate are located in the Kyle of Sutherland in the northeastern part of Scotland. Both 'Skibo' and 'Sutherland' are names whose roots are from Old Norse, the language spoken by the Vikings who began washing ashore regularly in the late ninth century. The word 'Skibo' fascinates etymologists, who are unable to agree on its original meaning. All agree that 'bo' is the Old Norse for 'land' or 'place,' but they argue whether 'ski' means 'ships' or 'peace' or 'fairy hill.'

Although the earliest version of Skibo seems to be lost in the mists of time, it was most likely some kind of fortified building erected by the Norsemen. The present-day castle was built by a bishop of the Roman Catholic Church. Andrew Carnegie, after making his fortune, bought it in 1898 to serve as his summer home. In 1980, his daughter, Margaret, donated Skibo to a trust that later sold the estate. It is presently being run as a luxury hotel.

CONTACT US

WEB PAGES
PDL Home: http://www.pdl.cmu.edu/
Please see our web pages at
http://www.pdl.cmu.edu/PEOPLE/
for further contact information.

**FACULTY**

Greg Ganger (pdl director)
412•268•1297
ganger@ece.cmu.edu
Anastassia Ailamaki
natassa@cs.cmu.edu
David Andersen
dga@cs.cmu.edu
Anthony Brockwell
abrock@stat.cmu.edu
Chuck Cranor
chuck@ece.cmu.edu
Christos Faloutsos
christos@cs.cmu.edu
Garth Gibson
garth@cs.cmu.edu
Seth Goldstein
seth@cs.cmu.edu
Mor Harchol-Balter
harchol@cs.cmu.edu
Todd Mowry
tcm@cs.cmu.edu
David O'Hallaron
droh@cs.cmu.edu
Adrian Perrig
adrian@ece.cmu.edu
Mike Reiter
reiter@cmu.edu
Mahadev Satyanarayanan
satya@cs.cmu.edu
Srinivasan Seshan
srini@cmu.edu
Dawn Song
dawnsong@ece.cmu.edu
Chenxi Wang
chenxi@ece.cmu.edu
Hui Zhang
hui.zhang@cs.cmu.edu

**POST DOCTORAL RESEARCHER**

Bianca Schroeder
bianca@cs.cmu.edu

**STAFF MEMBERS**

Bill Courtright 412•268•5485
(pdl executive director) wcourtright@cmu.edu
Karen Lindenfelser, 412•268•6716
(pdl business administrator) karen@ece.cmu.edu
Stan Bielski
Mike Bigrigg
Helen Borek-Conti
Joan Digney
Gregg Economou
Manish Prasad
Raja Sambasivan
Michael Stroucken
Terrence Wong

**GRADUATE STUDENTS**

Michael Abd-El-Malek
Mukesh Agrawal
Kinman Au
Shimin Chen
Ivan Dobrić
Kun Gao
Ben George
Nikos Hardavellas
James Hendricks
Rahul Iyer
Andrew Klosterman
Jure Leskovec
Julio López
Michael Mesnier
Jim Newsome
Amber Palekar
Jia-Yu Pan
Ippokratis Pandis
Stratos Papadomanolakis
Adam Pennington
Ginger Perng
Brandon Salmon
Minglong Shao
Shafeeq Sinnamohideen
Craig Soules
John Strunk
Jimeng Sun
Eno Thereska
Niraj Tolia
Tiankai Tu
Matthew Wachs
Shuheng Zhou

# FROM THE DIRECTOR'S CHAIR

We think the realness and experiences will be invaluable.

Planning for this deployment required finding machine room space, which is in short supply. This need, and complementary excitement around moving to a shared computing and storage utility, spawned the Data Center Observatory (DCO), which was conceived as a utility, an observatory, a testbed, and a showcase. The University administration has shown its support by allocating some of its most precious resource: space. Extensive effort has gone into designing a 2000 square foot machine room capable of supporting 40 high-density racks of equipment for computing, storage, and networking. Planning the room from an engineering standpoint has been an eye-opening experience, inducing new collaborations and new research foci. We have partnered with APC Corporation on power and cooling, and the DCO will demonstrate the latest in cooling technology and allow groundbreaking research in dynamic power+thermal management to reduce operational costs. We will need a lot of help to achieve the full Data Center Observatory vision, but it could be a huge step forward for data center research.

The new Computational Database Systems (CoDS) project, started last year, is making great strides in enabling a shift in scientific computing from ad hoc codes to more general-purpose database systems. Doing so simplifies scientists' programming tasks and allows them to benefit from decades of database optimization research and development. But, it does require changes to database systems, which have been optimized for business needs rather than scientific data explorations. Among other things (see the Computational Database Systems article in this PDL Packet), the CoDS research builds on the ongoing Fates database storage management work. A Fates-based database storage manager transparently exploits select knowledge of the underlying storage infrastructure to automatically achieve robust, tuned performance. The latest Fates ideas focus on exploiting disk characteristics to much more efficiently support access to multi-dimensional datasets on modern disk systems.

Of course, many other ongoing PDL projects are also producing cool results. For example, Connections is a context-enhanced semantic file system that uses temporal locality in file access patterns to expose inter-relationships that enable more effective file search. The self-securing devices project continues to explore intrusion survival features of augmenting devices with security functionality. The PASIS project has extended the scalable read/write storage protocols to general functions and shown them to be more fault-scalable than replicated state machines. This newsletter and the PDL website offer more details and additional research highlights.

I'm always overwhelmed by the accomplishments of the PDL students and staff, and it's a pleasure to work with them. As always, their accomplishments point at great things to come.

John and Andy visit the electrical power switching room in the basement of the CIC building to check out where the circuit breakers for the data center would be installed.

**October 2005**

- 13th Annual PDL Retreat and Workshop.
- Craig Soules will speak on "Connections: Using Context to Enhance File Search," and Jay Wylie will present "Fault-Scalable Byzantine Fault-Tolerant Services" at SOSP 2005 in Brighton, United Kingdom.
- Mike Abd-El-Malek will be speaking on "Lazy Verification in Fault-Tolerant Distributed Storage Systems" at SRDS 2005 in Orlando.

**September 2005**

- Stavros Harizopoulos successfully defended his dissertation on "Staged Database Systems" and has moved to Boston to take a post-doc position at MIT.
- Eno Thereska presented "Continuous Resource Monitoring for Self-predicting DBMS" and Raja Sambasivan spoke on "Replication Policies for Layered Clustering of NFS Servers" at MASCOTS 2005 in Atlanta, GA.
- Spiros Papadimitriou successfully defended his dissertation is moving to IBM Watson.
- Mengzhi Wang successfully defended her dissertation on "Performance Modeling of Storage Devices Using Machine Learning" and has just started working for Google New York.

**August 2005**

- Jure Leskovec, Jon Kleinberg and Christos Faloutsos received the best paper award for "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations" at KDD 2005, Chicago, IL.
- Jay Wylie successfully defended his dissertation on "A read/write protocol family for versatile storage infrastructures" and will be starting work at HP Labs in Palo Alto in November.
- Mike Abd-El-Malek received his M.S. for research on "Lazy verification in fault-tolerant distributed storage systems."
- Eno Thereska's paper "Towards self-predicting systems: What if you could ask 'what-if'?" was presented at the 3rd International Workshop on Self-adaptive and Autonomic Computing Systems in Copenhagen, Denmark.
- PDL researchers visited the Mellon Financial data center.
- Greg and Garth attended the High-End Computing I/O and File System Research Roadmapping workshop in Dallas, TX.
- John Bucy earned his M.S. degree for research on "Layout Characterization and Modeling for Modern Disk Drives" and joined Google.

**July 2005**

- Mellon Financial storage architects and administrators visited with PDL researchers at CMU.
- Eno Thereska proposed his Ph.D. research, titled "An instrumentation and performance querying framework for informed tuning in a self-managing system."

**June 2005**

- Bianca Schroeder successfully defended her dissertation on "Improving the Performance of Static and Dynamic Requests at a Busy Web Site" and is now a post-doctoral researcher with Garth Gibson.
- Niraj Tolia proposed his Ph.D. research, titled "Improving Conventional Client-Server Architectures Through Content Addressable Storage."
- Christos Faloutsos was a tutorial co-instructor at SIGMOD 2005 (Baltimore, Maryland) on "Research Issues in Protein Location Image Databases."
- Jay Wylie presented "Secure erasure-coded data" at the Second SNIA Security Summit in Pittsburgh, PA.
- Stavros Harizopoulos presented "QPipe: A Simultaneously Pipelined Relational Query Engine" at SIGMOD in Baltimore.
- APC joined the PDL Consortium

**May 2005**

- 7th annual Spring Industry Visit Day.
- Equal Logic joined the PDL Consortium.
- David Anderson presented "Improving Web Availability for Clients with MONET" at NSDI 2005 in Boston, MA.
- Minglong Shao interned at IBM Almaden Research Center, working in the Advanced Optimization, Advanced Database Solutions group.
- Mike Abd-El-Malek interned at Microsoft Research Cambridge, working on database query optimization.
- Shuheng Zhou has been visiting UC Berkeley since the end of the spring semester, working with Professor Satish Rao on designing approximation algorithms for network congestion minimization and Edge Disjoint paths problems.
- Angela Demke Brown successfully defended her dissertation on "Explicit Compiler-based Memory Management for Out-of-core Applications" and received the award for best SCS dissertation.
- PDL moved to the new CIC building.
- PDL co-hosted the SNIA Storage Security Summit (with CyLab).

*continued on page 21*

PDL members: present (Greg and Craig), past (Steve, now with Intel) and almost outta here. Good luck to Jay at HP.

### Comparison-Based File Server Verification

*Tan, Wong, Strunk & Ganger*

USENIX '05 Annual Technical Conference, April 10-15, 2005. Anaheim, CA.

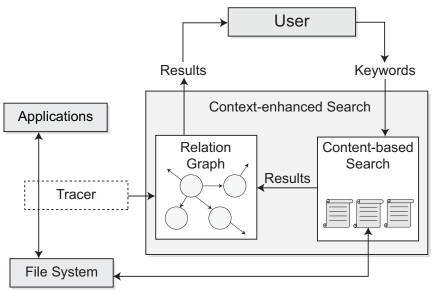
Comparison-based server verification involves testing a server by comparing its responses to those of a reference server. An intermediary, called a "server Tee," interposes between clients and the reference server, synchronizes the system-under-test (SUT) to match the reference server's state, duplicates each request for the SUT, and compares each pair of responses to identify any discrepancies. The result is a detailed view into any differences in how the SUT satisfies the client-server protocol specification, which can be invaluable in debugging servers, achieving bug compatibility, and isolating performance differences. This paper introduces, develops, and illustrates the use of comparison-based server verification. As a concrete example, it describes a NFSv3 Tee and reports on its use in identifying interesting differences in several production NFS servers and in debugging a prototype NFS server. These experiences confirm that comparison-based server verification can be a useful tool for server implementors.

### Connections: Using Context to Enhance File Search

*Soules & Ganger*

SOSP'05, October 23–26, 2005, Brighton, United Kingdom.

Connections is a file system search tool that combines traditional content-based search with context information gathered from user activity. By tracing file system calls, Connections can identify temporal relationships between files and use them to expand and reorder traditional content search results. Doing so improves both recall (reducing false positives) and precision (reducing false-negatives). For example, Connections improves the average recall (from 13% to 22%) and precision (from 23% to 29%) on the first ten results. When averaged across all recall levels, Connections improves precision from 17% to 28%. Connections provides these benefits with only modest increases in average query time (2 seconds), indexing time (23 seconds daily), and index size (under 1% of the user's data set).



Architecture of Connections. Both applications and the file system remain unchanged, as the only information required by Connections can be gathered either by a transparent tracing module or directly from existing file system interfaces.

### Challenges in Building a Two-Tiered Learning Architecture for Disk Layout

*Salmon, Thereska, Soules, Strunk & Ganger*

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-04-109, August, 2004.

Choosing the correct settings for large systems can be a daunting task. The performance of the system is often heavily dependent upon these settings, and the "correct" settings are often closely coupled with the workload. System designers usually resort to using a set of heuristic approaches that are known to work well in some cases. However, hand-combining these heuristics is painstaking and fragile. We propose a two-tiered architecture that makes this combination transparent and robust, and describe an application of the architecture to the problem of disk layout optimization. This two-tiered architecture consists of a set of independent heuristics, and an adaptive method of combining them. However, building such a system has proved to be more difficult than expected. Each heuristic depends heavily on decisions from other heuristics, making it difficult to break the problem into smaller pieces. This paper outlines our approaches and how they have worked, discusses the biggest challenges in building the system, and mentions additional possible solutions. Whether this problem is solvable is still open to debate, but the experiences reported provide a cautionary tale; system policy automation is complex and difficult.

### Towards Self-predicting Systems: What If You Could Ask "What-if"?

*Thereska, Narayanan & Ganger*

3rd International Workshop on Self-adaptive and Autonomic Computing Systems. Copenhagen, Denmark, August 2005. Supersedes Carnegie Mellon University Parallel Data Laboratory Technical Report CMU-PDL-05-101, February 2005.

Today, management and tuning questions are approached using if/then rules of thumb. This reactive approach requires expertise regarding of system behavior, making it difficult to deal with unforeseen uses of a system's resources and leading to system unpredictability and large system management overheads. We propose a "what... if..." approach that allows interactive exploration of the effects of system changes, thus converting complex tuning problem into simpler search problems. Through two concrete management problems, automating system upgrades and deciding on service migrations, we identify system design changes that enable a system to answer "what... if..." questions about itself.

*continued on page 6*

*continued from page 5*

### On Hierarchical Routing in Doubling Metrics

*Gupta, Maggs & Zhou*

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-04-106, December 2004.

We study the problem of routing in doubling metrics, and show how to perform hierarchical routing in such metrics with small stretch and compact routing tables (i.e., with a small amount of routing information stored at each vertex). We say that a metric $(X,d)$ has doubling dimension dim($X$) at most α if every set of diameter $D$ can be covered by $2^\alpha$ sets of diameter $D/2$. (A doubling metric is one whose doubling dimension $dim(X)$ is a constant.) For a connected graph $G$, whose shortest path distances $d_G$ induce the doubling metric (X, $d_G$), we show how to perform $(1 + \tau)$-stretch routing on $G$ for any $0 < \tau \leq 1$ with routing tables of size at most $(\alpha/\tau)^{O(\alpha)} log\Delta log\delta$ bits with only $(\alpha/\tau)^{O(\alpha)} log\ \Delta$ entries, where Δ is the diameter of $G$ and $\delta$ is the maximum degree of $G$. Hence the number of routing table entries is just $\tau$-$O(1)log\Delta$ for doubling metrics. These results extend and improve on those of Talwar (2004).

### QPipe: A Simultaneously Pipelined Relational Query Engine

*Harizopoulos, Shkapenyuk & Ailamaki*

SIGMOD 2005, June 14-16, 2005, Baltimore, Maryland, USA.

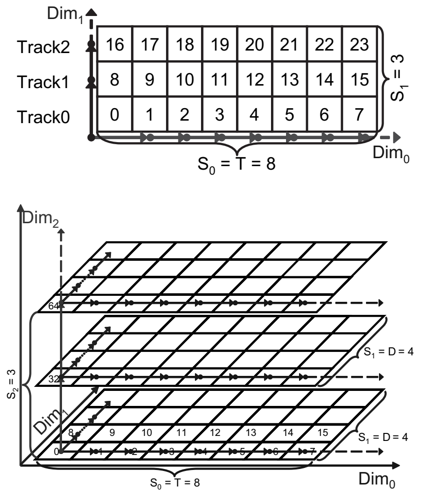
Relational DBMS typically execute concurrent queries independently by invoking a set of operator instances for each query. To exploit common data retrievals and computation in concurrent queries, researchers have proposed a wealth of techniques, ranging from buffering disk pages to constructing materialized views and optimizing multiple queries. The ideas proposed, however, are inherently limited by the query-centric philosophy of modern engine designs. Ideally, the query engine should proactively coordinate same-operator execution among concurrent queries, thereby exploiting common accesses to memory and disks as well as common intermediate result computation. This paper introduces on-demand simultaneous pipelining (OSP), a novel query evaluation paradigm for maximizing data and work sharing across concurrent queries at execution time. OSP enables proactive, dynamic operator sharing by pipelining the operator's output simultaneously to multiple parent nodes. This paper also introduces QPipe, a new operator-centric relational engine that effortlessly supports OSP. Each relational operator is encapsulated in a micro-engine serving query tasks from a queue, naturally exploiting all data and work sharing opportunities. Evaluation of QPipe built on top of BerkeleyDB shows that QPipe achieves a 2x speedup over a commercial DBMS when running a workload consisting of TPC-H queries.

### Multimap: Preserving Disk Locality for Multidimensional Datasets

*Shao, Schlosser, Papadomanolakis, Schindler, Ailamaki, Faloutsos & Ganger*

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-05-102, March 2005.

MultiMap is a new approach to mapping multidimensional datasets to the linear address space of storage systems. 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 overall I/O time by over 50% when compared to traditional naive layouts and by over 30% when compared to a Hilbert curve approach. For scans of the primary dimension, MultiMap and naive both provide almost two orders of magnitude higher throughput than the Hilbert curve approach.



Above, a 2-D dataset is mapped to disks. The first dimension is mapped to the track; the second dimension is mapped to the sequences of the first-a neighbors. Below, a 2-D dataset is mapped to disks. Each surface is a 2-D structure as in the 2-D map. The third dimension is mapped to the sequences of 4th-a neighbors.

### Fault-Scalable Byzantine Fault-Tolerant Services

*Abd-El-Malek, Ganger, Goodson, Reiter & Wylie*

SOSP'05, October 23-26, 2005, Brighton, United Kingdom.

A fault-scalable service can be configured to tolerate increasing numbers of faults without significant decreases in performance. The Query/Update (Q/U) protocol is a new tool that en-

*continued on page 7*

continued from page 6

ables construction of fault-scalable Byzantine fault-tolerant services. The optimistic quorum-based nature of the Q/U protocol allows it to provide better throughput and fault-scalability than replicated state machines using agreement-based protocols. A prototype service built using the Q/U protocol outperforms the same service built using a popular replicated state machine implementation at all system sizes in experiments that permit an optimistic execution. Moreover, the performance of the Q/U protocol decreases by only 36% as the number of Byzantine faults tolerated increases from one to five, whereas the performance of the replicated state machine decreases by 83%.

### Lazy Verification in Fault-Tolerant Distributed Storage Systems

*Abd-El-Malek, Ganger, Goodson, Reiter & Wylie*

24th IEEE Symposium on Reliable Distributed Systems (SRDS 2005), October 26-28, 2005, Orlando, Florida.

Verification of write operations is a crucial component of Byzantine fault-tolerant consistency protocols for storage. Lazy verification shifts this work out of the critical path of client operations. This shift enables the system to amortize verification effort over multiple operations, to perform verification during otherwise idle time, and to have only a subset of storage-nodes perform verification. This paper introduces lazy verification and describes implementation techniques for exploiting its potential. Measurements of lazy verification in a Byzantine fault-tolerant distributed storage system show that the cost of verification can be hidden from both the client read and write operation in workloads with idle periods. Furthermore, in workloads without idle periods, lazy verification amortizes the cost of verification over many versions and so provides a factor of four higher write bandwidth when compared to performing verification during each write operation.

### Replication Policies for Layered Clustering of NFS Servers

*Sambasivan, Klosterman & Ganger*

13th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), September 27-29, Atlanta, GA.

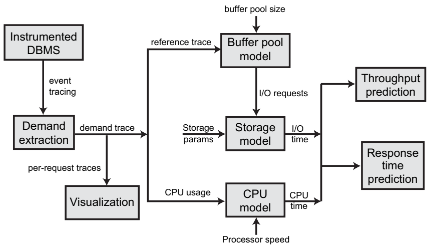
Layered clustering offers cluster-like load balancing for unmodified NFS or CIFS servers. Read requests sent to a busy server can be offloaded to other servers holding replicas of the accessed files. This paper explores a key design question for this approach: which files should be replicated? We find that the popular policy of replicating readonly files offers little benefit. A policy that replicates readonly portions of read-mostly files, however, implicitly coordinates with client cache invalidations and thereby allows almost all read operations to be offloaded. In a read-heavy trace, 75% of all operations and 52% of all data transfers can be offloaded.

### Continuous Resource Monitoring for Self-predicting DBMS

*Narayanan, Thereska & Ailamaki*

13th Annual Meeting of the IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2005), Atlanta, GA, September 27-29, 2005.

Administration tasks increasingly dominate the total cost of ownership of database management systems. A key task, and a very difficult one for an administrator, is to justify upgrades of CPU, memory and storage resources with quantitative predictions of the expected improvement in workload performance. Current database systems are not designed with such prediction in mind and hence offer only limited help to the administrator. This paper proposes changes to database system design that enable a Resource Advisor to answer "what-if" questions about resource upgrades. A prototype Resource Advisor built to work with a commercial DBMS shows the efficacy of our approach in predicting the effect of upgrading a key resource—buffer pool size—on OLTP workloads in a highly concurrent system.



Resource Advisor components.

### Modeling the Relative Fitness of Storage Devices

*Mesnier, Wachs & Ganger*

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-05-106, August, 2005.

Relative fitness modeling is a new approach for predicting the performance and resource utilization of a workload when running on a particular storage device. In contrast with conventional device models, which expect device independent workload characteristics as input, a relative fitness model makes predictions based on characteristics measured on a specific other device. As such, relative fitness models explicitly account for the workload changes that almost always result from moving a workload across storage devices—for example, higher I/O performance

continued on page 16

**October 2005**

### Many Completed Ph.D.s

Many PDL students have completed their Ph.D.s and moved on to new vistas in the past year. Greg had three students graduate. Jay Wylie defended in August and is going to HP Labs in Palo Alto. Garth Goodson finished last summer and moved to California to join Network Appliance, and Chris Lumb defended last November and is now with DataDomain. As well, John Bucy and Mike Abd-El-Malek have completed their MS degrees. John has moved on to Google and Mike is continuing his research in pursuit of a Ph.D. at CMU. David Petrou, co-advised by Greg and Garth, finished last December.

Spiros Papadimitriou, advised by Christos Faloutsos completed his Ph.D. work in September and has been with IBM Watson since the beginning of October.

Natassa's first student to complete his Ph.D., Stavros Harizopoulos, defended on September 19, and is joining Mike Stonebraker's group at MIT as a post-doctoral researcher. Mengzhi Wang defended on September 29 and is now at Google's New York office.

Natassa and Todd co-advised two graduating students: Chris Colohan, who has gone to Google and Shimin Chen, who will begin at Intel Research Pittsburgh in October. Angela Demke Brown, who is an Assistant Professor at the University of Toronto defended her Ph.D. research in April. See below for details on the awards garnered by her research.

**August 2005**

### Mike Bigrigg forms Spinoff

Researcher Michael Bigrigg, a project scientist in the Institute for Complex Engineered Systems, recently started a company that manufactures a sensor to detect and mitigate temperature problems in computer hard drives. The Pittsburgh-based company, Pervasive Sensors Inc., is producing Critter™, a computer-based temperature sensor that monitors and helps regulate adverse environmental conditions. The $21 device attaches to a desktop computer's game port and can be installed in minutes, requiring no special knowledge of computer software. The spinoff company is a result of research performed at the university over the past two years. For more information, see the CMU press release.

—with info from CMU's 8 1/2 x 11 News

**August 2005**

### Angela Demke Brown Co-winner of Best SCS Dissertation and Nominated for Best ACM Dissertation

Carnegie Mellon's two nominees for the ACM Doctoral Dissertation Award for 2004-05 have been announced. Angela Demke Brown has been nominated for her work on Explicit Compiler-based Memory Management for Out-of-core Applications (Todd Mowry, Advisor), following her selection as a co-winner of the 2005 SCS Dissertation Award. Also selected was Sanjit Seshia for Adaptive Eager Boolean Encoding for Arithmetic Reasoning in Verification (Randy Bryant, Advisor). Angela will share a cash prize and will be honored in the SCS Distinguished Lecturer series. These nominees were chosen by the Doctoral Dissertation Award Committee, chaired by Stephen Brookes. ACM nominees participate in the ACM evaluation process, representing Carnegie Mellon and competing against other nominees from universities throughout the United States. Congratulations Angela!

**March 2005**

### WVCIII!

It is with great pride that Bill and Mireille Courtright share the news of the birth of William Vance Courtright III. William was born on Tuesday, March 22, 2005 at 2:30 am. He was born 7 lbs. 7 oz. and 21.5 inches in length. All our best to the Courtright family!

**March 2005**

### Matthew Wachs awarded NDSEG Fellowship

Congratulations to Matthew Wachs, who has been selected to receive a 2005-2006 National Defense Science and Engineering Graduate (NDSEG) Fellowship. This fellowship is sponsored by the Department of Defense through the Air Force Office of Scientific Research (AFOSR), the Office of Naval Research (ONR), the Army Research Office (ARO), and the High Performance Computing Modernization Program, and is administered by the American Society for Engineering Education (ASEE).

**March 2005**

### Dawn Song Receives NSF CAREER AWARD

Dawn Song, assistant professor of Electrical and Computer Engineering and Computer Science, has received a National Science Foundation CAREER Award for her research proposal, "Toward Exterminating Large

*continued on page 9*

*continued from page 8*

Scale Internet Attacks." The award "recognizes and supports the early career-development activities of those teacher-scholars who are most likely to become the academic leaders of the 21st century."

—with info from CMU's 8 1/2 x 11 News

**February 2005**

### David Andersen Receives Award for Outstanding Ph.D. Thesis

Congratulations to David for being selected as the recipient of the 2005 MIT EECS George M. Sprowls Award for outstanding Ph.D. thesis. Dave's work covered "Improving End-to-End Availability Using Overlay Networks." His thesis explores internet service failure and methods to improve service using three complementary overlay networks—networks created dynamically between a group of cooperating Internet hosts—to determine whether the Internet path between two hosts is working on an end-to-end basis, exploiting the considerable redundancy available in the underlying Internet to find working paths, and to guard against denial-of-service attacks.

**January 2005**

### Anastassia Ailamaki selected as a Sloan Research Fellow

We are extremely pleased to announce that Anastassia Ailamaki has been selected as a winner of a Sloan Research Fellowship. 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 116 fellowships are awarded annually in seven fields: chemistry, computational and evolutionary molecular biology, computer science, economics, mathematics, neuroscience, and physics.

**January 2005**

### Welcome Evan!

Joan and Bruce Digney are thrilled to announce the arrival of their son Evan Bruce. He arrived at 2:51 am on January 23, weighing 8 lbs. 6.5 oz and measuring 20.25 inches. Since then, he has grown considerably and is a delight to his Mom and Dad.

**December 2004**

### Garth and Vianey Marry!

Congratulations to the Goodsons! Garth and Vianey were married on December 19th, 2004 in Cuernavaca, Mexico at the Hacienda de Cortes (former home to the son of Cortes).

**December 2004**

### Outstanding Researchers

Prof. David O'Hallaron and Research Engineer Volkan Akcelik have been awarded the Outstanding Research Award for their work on the Quake Project by the College of Engineering in its CIT Faculty Awards for 2004–2005. The Quake project is joint effort by the Dept. of Civil and Environmental Engineering and the School of Computer Science at Carnegie Mellon University. Its goal is to develop the capability for predicting, by computer simulation, the ground motion of large basins during strong earthquakes, and to use this capability to study the seismic response of the Greater Los Angeles Basin.

—with info from CMU's 8 1/2 x 11 News

**November 2004**

### Pittsburgh – The Epicenter of Storage Innovation

A reception highlighting Pittsburgh as the Epicenter of Storage Innovation was held on November 11, 2004 as a part of Supercomputing 2004, hosted by Pittsburgh's Storage Innovators, include the Data Storage Systems Center (DSSC) at CMU, Intel Research Pittsburgh, Network Appliance, Panasas, the PDL at CMU, Pittsburgh Digital Greenhouse, Pittsburgh Supercomputing Center and Seagate Research.

SC04 marks the first year that storage has gotten explicit recognition in high-performance computing, though Pittsburgh has long been a leader in storage innovations. Jim Morris, former Dean of the School of Computer Science and current Dean of CMU's

The PDL, in collaboration with the DSSC, participated in the Epicenter of Storage Innovation Expo at Supercomputing in Pittsburgh in November.

*continued on page 21*

*continued from page 1*

as they are refined, in other environments as well. Beyond understanding costs, deploying proposed solutions in a real environment will allow us to measure how well they work in practice when dealing with the oddities of a real data center environment. Although anecdotal, such real experiences will complement lab experimentation well in proving ideas.

Carnegie Mellon's administration has gotten behind the DCO effort, recognizing its value for CMU (reduced operational costs) and the research community at large (knowledge). They have allocated for us their most precious resource—space—and are helping with construction costs for the data center observatory space. Their hope, of course, is that the result is substantial long-term savings in human adminsitration overheads, equipment acquisitions (by sharing hardware resources), total machine room space demands, power bills, and chilled water demands. By enabling CMU's researchers to attack these challenges, they are investing in a better future.

### Logistics (Size, Location, Scale)

The DCO will abut the main hallway of the main floor of the Collaborative Innovation Center (CIC), which is the newest building on the CMU campus. It is approximately 2000 square feet in size, and an adjacent 1100 square foot room will provide space for administrator cubicles, construction/testing benches, storage, and unpacking. The location was chosen to maximize visibility, such that the DCO is a showcase. There will be a windowed wall for walk-by tours and general observers and a display that continuously tells the research story and shows measurement data (thermal graphs, utilizations, etc.).

The room has been designed, in partnership with APC Corporation, to be fully instrumented and adaptively controlled. Physical sensors will measure all of the environmental aspects, including electrical, chilled water, temperature, and smoke. The DCO will have a private chilled water loop, with variable rate control, that is connected to the campus chilled water supply via heat exchangers. The cooling equipment has variable speed fans and other adaptive control options. These latter features enable adaptive power and cooling management (see below).

The engineering plans allow for a peak capacity of 40 20KW racks of computing equipment, plus the additional power needed for in-room cooling and overcoming power distribution losses. The plans call for equipment to be staged into the room, over a number of years, as demand for the shared computation and storage resources grows. (The demand will grow as research groups on campus transition, one by one, from private clusters to the shared infrastructure.) Multiple network fibres will connect the DCO to the primary campus backbone.

As of October 2005, engineering plans have been completed and put out for construction bids. We expect to begin installing equipment into the DCO in January 2006.

### Partnering with APC on Power & Cooling

APC (American Power Conversion) Corporation has partnered with us on the physical design of the power and cooling infrastructure, and its monitoring and control. Their novel In-Row™ cooling and Hot Aisle containment™ concepts are enabling remarkable equipment densities, enabling us to maximize usage of the DCO space while accommodating the high-density racks of today's and tomorrow's data center and allowing easy incremental deployment. APC's technologies also provide for thermal control with lower power requirements than conventional schemes. Additionally, APC is helping us to instrument and adaptively control all aspects of thermal and power management.

Greg Ganger and Dave Roden of APC look at cooling mechanisms in a high density enclosure (HDE). It is two rows of racks enclosed with doors on the end and a roof on top such that the hot aisle is inside the enclosed area. Precision air handlers in the HDE draw the hot air from the inside (hot aisle) and return it to the room at room temperature.

We plan to exploit the detailed instrumentation and adaptive control to minimize wasted energy. Most clusters and storage servers deployed on campus, and in data centers in general, run continuously despite their bursty usage. Although heavily utilized before paper submission deadlines, for example, university hardware resources are very lightly utilized at other times. Unfortunately, measurements of existing clusters indicate that idle machines still use approximately two-thirds of the power of maximully-loaded machines. Much energy is wasted.

As one area of DCO research (and benefit for CMU), we will be developing integrated approaches to resource assignment and energy conservation. Computation jobs and storage activity will be distributed among the DCO machines, and unutilized machines will be turned off. As research into use of service-level objectives (SLOs) matures and becomes part of DCO operation, more informed decisions about when to turn machines on and off will be possible. As machines are turned off, A/C units can also be turned down to lower power levels, saving further energy. In fact, careful choice of *which* machines to turn off can enhance power savings, as harder working cooling units can be offloaded to achieve superlinear benefits.

*continued on page 11*

## DATA CENTER OBSERVATORY

*continued from page 10*

### Self-* Storage

The initial seed that grew into the DCO vision was PDL's Self-* Storage project. We are exploring new storage architectures that integrate automated management functions and simplify adminsitration of storage systems. "Self-*" storage systems are self-configuring, self-organizing, self-tuning, self-healing, self-managing systems of storage bricks (small-scale storage servers). Thus, each step taken towards the ideal of self-* storage should simplify storage adminstration, reduce system cost, increase system robustness, and simplify system construction. To gain experience with and evaluate our ideas, we have long planned on deploying a large-scale storage service (100s of TBs), offering storage to CMU research groups and services. We are convinced that such deployment and maintenance are necessary to evaluate the ability of new architectures/technologies to simplify administration for the system scales and workload mixes that traditionally present difficulties. The DCO grew from the initial Self-* Storage planning, as it became clear that (1) machine room space would be an issue on campus, and (2) that dealing with the storage in isolation left too much outside the scope and created major performance difficulties.

The early stages of the self-* storage project have involved large amounts of design and infrastructure building. The foci during this period has been on enabling high levels of fault-tolerance, versatility, and instrumentation. Together, these can provide a foundation for the aggregation and virtualization, automated decision making, feedback control, diagnosis, and repair mechanisms needed to achieve self-*-ness. Research on such challenges, as well as scalability of a shared cluster-based storage infrastructure, is being pursued along many tracks.

As a platform for such research, and our planned deployment, we have built our first prototype. Ursa Minor is a cluster-based storage system that allows data-specific selection of, and on-line changes to, encoding schemes and fault models. Thus, different data

*continued on page 19*

## NEW PDL FACES

### Bill Courtright

We are very pleased to welcome Bill Courtright back to the PDL as Executive Director as of December, 2004. Bill initially joined the PDL as its executive director in 1998, but went on leave in 1999 to co-found Panasas, a storage systems company with headquarters in Fremont, CA. He initially served as the company's Chief Operating Officer, responsible for product development, financial and shareholder operations, and human resources. As the company matured, he focused his efforts on engineering program management and intellectual property management.

Bill is looking forward to working with Greg and Karen to ensure that consortium members get the most out of their partnership with the PDL. He also hopes to share his industry and entrepreneurial experiences with the students. The PDL's current focus on Self-* Storage is an ideal venue for his expertise in moving projects from development and into deployment, and a big part of this effort will be overseeing the design and construction of the Data Center Observatory, along with many other technical contributions to the storage management portion of the DCO project.

Bill received his B.S. in Electrical Engineering from the University of Kansas in 1986. He then became the lead design engineer for five hardware products at NCR's Storage Systems Division from 1986 to 1992, with responsibilities from concept through introduction to manufacturing and customer acceptance. During this period he earned an MS in Computer Engineering from the National Technological University.

He was then granted a sponsored academic leave between 1993 and 1995 to attend graduate school at Carnegie Mellon University. He was one of the founding members of the PDL, which held its first workshop in 1993—he is definitely a PDL member of long standing! Bill was awarded a Ph.D. in Electrical & Computer Engineering for his work in the use of transactional mechanisms in the design and implementation of redundant disk array software. This work led to the production of RAIDframe, a RAID implementation that is deployed in industrial and academic settings, including availability in NetBSD.

After returning to NCR (which had since become Symbios Logic), Bill worked in the strategic marketing organization, contributing to new business development, followed by a year in the server and storage architecture organization. While in this latter position, he played a significant role in the development of the architecture and implementation of a next-generation storage management system. Between 1996 and 1998, Bill was also responsible for all patent activity in the Storage Systems Division of Symbios Logic; he is a co-inventor of eight patents in the field of storage systems.

# COMPUTATIONAL DATABASE SYSTEMS

*Anastassia Ailamaki & David O'Hallaron*

Dramatic increases in computing power and storage capacity have allowed scientists to build simulations that model nature in more detail than ever. However, massive computations often require massive input and output datasets, and the size of these datasets is rapidly outpacing scientists' ability to manipulate and use them.

For the past 10 years, the Carnegie Mellon Quake group has been building computer models that predict the motion of the ground during strong earthquakes. The goal of the Quake project is to predict, by computer simulation, the ground motion of large basins during strong earthquakes, and to use the results to study the seismic response of the Greater Los Angeles Basin. It does not try to predict earthquakes; rather, it strives to answer other questions: if a large earthquake strikes Los Angeles, which regions will be worst stricken? Which seismic frequencies will be amplified most by the soil? Answers will help architects design buildings whose resonant frequencies are least excited during an earthquake.

A geological model of soil densities may be repeatedly queried by a mesh generator to build an unstructured mesh that consists of a collection of hexahedral (brick-like) elements, and a collection of nodes, (the corners of the elements). Regions with softer soils are represented by smaller elements, and regions with harder soils get larger ones. Hexmeshes are attractive because they combine the multi-scale resolution of arbitrarily unstructured meshes with the simpler topology of regularly structured grids. A solver code simulates the propagation of seismic waves through the earth by approximating the solution to the wave equation at each of the mesh nodes. During each time step, the solver computes an estimate of each node velocity in three directions and writes the resulting floating-point numbers to disk. The result is a four-dimensional spatio-temporal earthquake dataset that describes the velocity response of the ground. For a billion-node mesh, each timestep requires about 12 GB. Simulating 60 seconds of shaking requires about 20K timesteps. Thus, storing all of the information from one simulation requires ~25 TB. The earthquake dataset may then be queried in different ways to support different types of analyses. In 1993, the largest simulation code required an input unstructured finite element mesh with 50K nodes (1.5 MB) and produced a relatively small 500 MB output dataset. By 2003, the largest LAB simulation required a mesh with 1.37B nodes (45 GB) and generated an output dataset that was over one TB.

With such massive datasets involved, previously routine activities, such as generating unstructured meshes to define earthquake source models and partitioning meshes for parallel computing, become challenging tasks. This is primarily because unstructured datasets require complex pointer-based structures to represent them and because the massive amounts of main memory required means these dataset manipulations can no longer be conducted by scientists on their desktop computers.

Point queries are aggregated in both space and time. Engineering analyses require time-varying queries where the spatial coordinates are constant and time varies, while visualizations require space-varying queries where the spatial coordinates vary and time is constant. One problem faced by the aggregate queries is that data layouts favoring time-varying queries tend to punish space-varying queries, and vice versa. A key research task is to develop database and storage structures that will allow support of both types of queries. Another important task is to develop compression techniques that will still allow fast queries over the compressed datasets.

At the first glance, a ready solution would seem to be to take an existing relational database management system (RDBMS), embed SQL commands in simulation codes and then run the codes on the database. Unfortunately, conventional databases are designed and optimized for business applications, rather than physical simulations, which are much more dynamic, being produced, updated and removed on the fly at a high rate. If every data access operation was implemented with a standard SQL command (select, insert, update or delete), the overhead would be far too large for such programs to handle.

## Computational Data Systems

How can both performance and functionality exist side by side? A new system has been developed that incorporates computational data access patterns of physical simulations into the design of the underlying database structures, and exports a specialized set of tightly-coupled and highly-optimized functions to interact with

*continued on page 13*

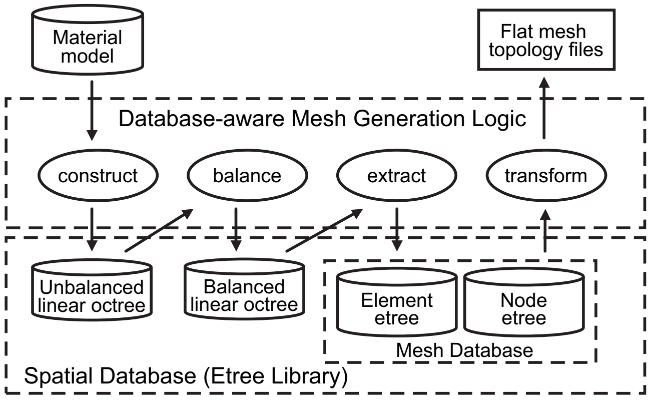


Figure 1: Structure and Workflow of the Weaver system.

continued from page 12

the databases. To differentiate such systems from conventional RDBMSs, we refer to them as Computational Database Systems. Computational data systems integrate scientific datasets stored as spatial databases with specialized, tightly-coupled, and highly-optimized functions that query and manipulate these databases. Doing so simultaneously simplifies domain-specific programming (since datasets are now in a form that can be queried generally), and greatly increases the scale of problems that can be studied efficiently.

This is appealing on several fronts. First, because every step of the simulation process works by operating on databases, the sizes of the models are only limited by the amount of available disk space, rather than the amount of available memory. Second, by integrating physical simulations with database systems, we are able to build on decades of previous database research, borrowing ideas such as B-tree/R-tree indexing, linear quadtrees, space-filling curves, cache-aware data layout, etc., augmenting these with new algorithms and techniques. Third, by managing data on behalf of simulation codes, it is possible to organize the data in such a way that best exploits locality. At runtime, such locality can be carried up through the memory hierarchy, improving performance. In addition, by understanding how simulation codes are going to access the data, it becomes possible to implement effective data prefetching techniques without modifying application code.

## Weaver

Driven by the CMU Quake project to generate higher resolution unstructured meshes, a prototype computational data system named Weaver has been developed capable of generating massive, queryable unstructured octree-based hexahedral meshes on desktop systems. Among many different types of meshes, octree-based hexahedral meshes lie in between the extremes of arbitrarily unstructured meshes and regularly structured meshes. They provide a compromise between modeling power and simplicity. On the one hand, they are able to subdivide an octant to resolve local heterogeneity and provide multi-scale resolution as do other unstructured meshes. On the other hand, they produce only one primitive shape for all elements. The recursive process of subdivision leads to a relatively structured placement of mesh nodes, similar to many regularly structured meshes. The core idea of the Weaver system is that it stores and indexes (partially generated) hexahedral meshes using spatial database structures, and implements the mesh generation process by querying and manipulating the database. A limitation of the Weaver system is that it is not an on-line solution-adaptive mesh generator that can dynamically adjust a mesh structure while a solver is working on the mesh.

Conceptually, the Weaver system consists of two parts: a spatial database and mesh generation logic. The spatial database manages unstructured mesh data (elements and nodes) on disk and in memory data. The mesh generation logic implements different mesh generation steps by exploiting the characteristics of the database structure in order to reduce disk I/O and improve the running time (time-complexity). The Weaver mesh generation logic consists of the following closely related steps, as shown in Figure 1. The construct step builds an indexed linear octree on disk. The sizes of the octants are determined by an application, such as by the density of the material they enclose. The balance step recursively subdivides octants as necessary to enforce the 2-to-1 constraint. The extract step uses the balanced linear octree as a template to generate a queryable mesh database that consists of an element etree and a node etree. The transform step queries the data in the mesh database and generates a flat mesh topology file that establishes the element-node connectivity relationship.

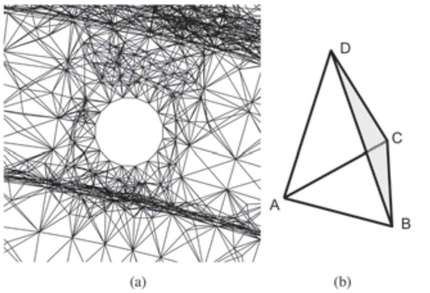


Figure 2: (a) Part of a 3-dimensional tetrahedral mesh with complex geometry, and (b) A single tetrahedral element with nodes A, B, C, D.

When used to model heterogeneous geological structures such as sedimentary basins where material properties vary significantly throughout the domain, multi-resolution unstructured hexahedral meshes allow a tremendous reduction (approx. three orders of magnitude) in the number of mesh nodes (compared to uniform meshes), because element sizes can adapt locally to the highly-variable wavelength of propagating seismic waves. The Weaver system is capable of generating extremely large meshes on a desktop system in a reasonable amount of time. For example, a 2 Hz mesh, with 1.37B nodes, involves creating a mesh database of size 45GB and flat topology files of size 80.5GB. If all mesh data structures had been built in main memory, more than 300 GB memory would have been used (on an Alpha system with 8-byte pointers). This implies that the Weaver system is exploiting locality efficiently and is processing data mostly from within memory. Otherwise, large fluctuations due to disk I/O would result as the problem size increased.

## Directed Local Search

Researchers at CMU have developed another novel query processing

continued on page 24

## DISSERTATION ABSTRACT:

### Explicit Compiler-based Memory Management for Out-of-core Applications

*Angela Demke Brown, CS*

*Carnegie Mellon University Ph.D Dissertation CMU-CS-05-140, May 2005.*

For a large class of scientific computing applications, the continuing growth in physical memory capacity cannot be expected to eliminate the need to perform I/O throughout their executions. For these out-of-core applications, the large and widening gap between processor performance and disk latency is a major concern. Current operating systems deliver poor performance when an application's working set does not fit in main memory. As a result, programmers who wish to solve these out-of-core problems efficiently are typically faced with the onerous task of rewriting their application to use explicit I/O operations (e.g., read/write). In many cases, the end result is that the size of physical memory determines the size of problem that can be solved.

In this dissertation, we propose and evaluate a fully-automatic technique which liberates the programmer from this task, provides high performance, and requires only minimal changes to current operating systems. In our scheme, the compiler provides the crucial information on future access patterns without burdening the programmer, the operating system supports non-binding prefetch and release hints for managing I/O in a virtual memory system, and the operating system cooperates with a run-time layer to accelerate performance by adapting to dynamic behavior and minimizing prefetch overhead. This approach maintains the abstraction of unlimited virtual memory for the programmer, gives the compiler the flexibility to aggressively insert prefetches ahead of references, and gives the operating system the flexibility to arbitrate between the competing resource demands of multiple applications.

We implemented our compiler analysis within the SUIF compiler, and used it to target implementations of our run-time and operating system support on both research and commercial systems (HURRICANE and IRIX 6.5, respectively). Our experimental results show large performance gains for out-of-core scientific applications on both systems: more than 50% of the I/O stall time has been eliminated in most cases, thus translating into overall speedups of roughly twofold in many cases. Our initial experiments motivated a new compiler scheduling algorithm that is capable of tolerating the large and variable latencies that are common for disk accesses, in the presence of multiply-nested loops with unknown bounds. On our current experimental systems, many of our benchmark applications remain I/O bound, however, we show that the new scheduling algorithms are able to substantially improve performance in some cases, reducing execution time by an additional 36% in the best case. We further show that the new algorithms should enable applications to make more effective use of higher-bandwidth disk systems that will be available in the future.

## DISSERTATION ABSTRACT:

### Cluster Scheduling for Explicitly-Speculative Tasks

*David Petrou, ECE*

*Carnegie Mellon University Ph.D. Dissertation CMU-PDL-04-112, December 2004.*

A process scheduler on a shared cluster, grid, or supercomputer that is informed which submitted tasks are possibly unneeded speculative tasks can use this knowledge to better support increasingly prevalent user work habits, lowering user-visible response time, lowering user costs, and increasing resource provider revenue.

Large-scale computing often consists of many speculative tasks (tasks that may be canceled) to test hypotheses, search for insights, and review potentially finished products. For example, speculative tasks are issued by bioinformaticists comparing DNA sequences, computer graphics artists rendering scenes, and computer researchers studying caching. This behavior—exploratory searches and parameter studies, made more common by the cost effectiveness of cluster computing—on existing schedulers without speculative task support results in a mismatch of goals and suboptimal scheduling. Users wish to reduce their time waiting for needed task output and the amount they will be charged for unneeded speculation, making it unclear to the user how many speculative tasks they should submit.

This thesis introduces 'batchactive' scheduling (combining batch and interactive characteristics) to exploit the inherent speculation in common application scenarios. With a batchactive scheduler, users submit explicitly

*continued on page 15*

Mike Mesnier talks on "Modeling the Relative Performance of Self-* Storage Bricks" at the 2004 PDL Retreat.

continued from page 14

labeled batches of speculative tasks exploring ambitious lines of inquiry, and users interactively request task outputs when these outputs are found to be needed. After receiving and considering an output for some time, a user decides whether to request more outputs, cancel tasks, or disclose new speculative tasks. Users are encouraged to disclose more computation because batchactive scheduling intelligently prioritizes among speculative and non-speculative tasks, providing good wait-time-based metrics, and because batchactive scheduling employs an incentive pricing mechanism which charges for only requested task outputs (i.e., unneeded speculative tasks are not charged), providing better cost-based metrics for users. These aspects can lead to higher billed server utilization, encouraging batchactive adoption by resource providers organized as either cost- or profit-centers.

Not all tasks are equal—only tasks whose outputs users eventually desire matter—leading me to introduce the ‘visible response time’ metric which accrues for each task in a batch of potentially speculative tasks when the user needs its output, not when the entire batch was submitted, and the batchactive pricing mechanism of charging for only needed tasks, which encourages users to disclosure future work while remaining resilient to abuse. I argue that the existence of user think times, user away periods, and server idle time makes batchactive scheduling applicable to today’s systems.

I study the behavior of speculative and non-speculative scheduling using a highly-parameterizable discrete-event simulator of user and task behavior based on important application scenarios. I contribute this simulator to the community for further scheduling research.

For example, over a broad range of realistic simulated user behavior and task characteristics, I show that under a batchactive scheduler visible response time is improved by at least a factor of two for 20% of the Simulations. A batchactive scheduler which favors users who historically have desired a greater fraction of tasks that they speculatively disclosed provides additional performance and is resilient to a denial-of-service. Another result is that visible response time can be improved while increasing the throughput of tasks whose outputs were Desired. Under some situations, user costs decrease while server revenue increases. A related result is that more users can be supported and greater server revenue generated while achieving the same mean visible response time. A comparison against an impractical batchactive scheduler shows that the easily implementable two tiered batchactive schedulers, out of all batchactive schedulers, provide most of the potential performance gains. Finally, I demonstrate deployment feasibility by describing how to integrate a batchactive scheduler with a popular clustering system.

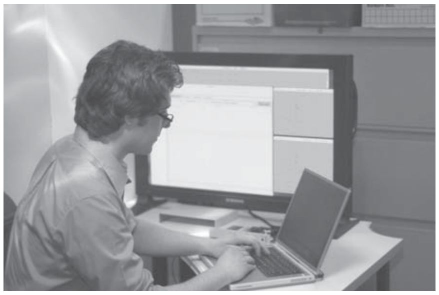
## DISSERTATION ABSTRACT:

### Improving the Performance of Static and Dynamic Requests at a Busy Web Site

*Bianca Schroeder, CS*

*Carnegie Mellon University Ph.D. Dissertation, June 21, 2005.*

Running a high-volume web site is a challenging task. Web traffic is bursty with peak request rates rising far above average rates and likely phenomena like flash crowds and hot spots. Yet web users are very demanding: they expect Web sites to be quickly accessible from around the world 24 hours a day, 7 days a week. Recent studies show that even a short period of slowdown or service interruption can have severe effects: it not only sends customers to the "just a click-away" competitor; it also reflects negatively on the corporate image as a whole.



Stan Bielski demonstrates “Coordination and Configuration of Self-Securing Network Interfaces” at the 2005 PDL Spring Industry Visit Day.

The broad goal of this thesis is to improve the user-perceived performance of both static and dynamic requests at a busy web site. By static requests we mean requests of the form “GET me a file”. By dynamic requests we mean those which require the web server to create the requested information on the fly, by accessing a database backend. The approach taken in this thesis does not require buying more hardware or any other costly system upgrades. The main idea is to schedule the existing resources better among requests so as to either improve overall mean performance or improve the performance of a small subset of high-priority requests.

The first part of the thesis presents a very simple solution for improving overall mean response times for static web workloads by favoring those requests that are quick, or have small remaining processing requirements in accordance with the SRPT scheduling policy. Our policy is particularly effective during transient overload. The second part of this thesis focuses on the more complex dynamic web requests. We propose and implement various approaches for providing differentiated levels of service for database-driven dynamic requests. We propose and evaluate algorithms for both scheduling of database internal resources and scheduling outside the DBMS through an external front-end. In the third part of the thesis we study the effect of the experimental

continued on page 18

continued from page 7

usually leads to faster application execution which results in higher I/O rates. Further, relative fitness models allow service observations (e.g., performance and resource utilizations) from the measured device to be used in making predictions on the modeled device—such observations often provide more predictability than basic workload characteristics. Overall, we find that relative fitness models reduce prediction error by over 60% on average when compared to conventional modeling techniques.

### Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations

*Leskovec, Kleinberg & Faloutsos*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2005), Chicago, IL, USA, 2005.

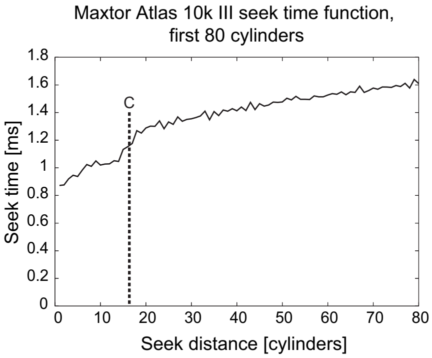
How do real graphs evolve over time? What are "normal" growth patterns in social, technological, and information networks? Many studies have discovered patterns in *static graphs*, identifying properties in a single snapshot of a large network, or in a very small number of snapshots; these include heavy tails for in- and out-degree distributions, communities, small-world phenomena, and others. However, given the lack of information about network evolution over long periods, it has been hard to convert these findings into statements about trends over time. Here we study a wide range of real graphs, and we observe some surprising phenomena. First, most of these graphs densify over time, with the number of edges growing superlinearly in the number of nodes. Second, the average distance between nodes often shrinks over time, in contrast to the conventional wisdom that such distance parameters should increase slowly as a function of the number of nodes (like *O*(log *n*) or *O*(log(log *n*)). Existing graph generation models do not exhibit these types of behavior, even at a qualitative level. We provide a new graph generator, based on a "forest fire" spreading process, that has a simple, intuitive justification, requires very few parameters (like the "flammability" of nodes), and produces graphs exhibiting the full range of properties observed both in prior work and in the present study.

### On Multidimensional Data and Modern Disks

*Schlosser, Schindler, Shao, Papadomanolakis, Ailamaki, Faloutsos & Ganger*

Proceedings of the 4th USENIX Conference on File and Storage Technologies (FAST '05). San Francisco, CA. December 13-16, 2005.

With the well-ingrained notion that disks can efficiently access only one dimensional data, current approaches for mapping multidimensional data to disk blocks either allow efficient accesses in only one dimension, trading off the efficiency of accesses in other dimensions, or equally penalize access to all dimensions. Yet, existing technology and functions readily available inside disk firmware can identify non-contiguous logical blocks that preserve spatial locality of multidimensional datasets. These blocks, which span on the order of a hundred adjacent tracks, can be accessed with minimal positioning cost. This paper details these technologies, analyzes their trends, and shows how they can be exposed to applications while maintaining existing abstractions. The described approach can achieve the best possible access efficiency afforded by the disk technologies: sequential access along primary dimension and access with minimal positioning cost for all other dimensions. Experimental evaluation of a prototype implementation demonstrates a reduction of the overall I/O time between 30% and 50% for multidimensional data queries when compared to existing approaches.



Measured seek curve of Maxtor Atlas10k3, first 80 cylinders. The transition from settle-dominated to seek-dominated positioning time occurs at a distance of twelve cylinders.

### Ursa Minor: Versatile Cluster-based Storage

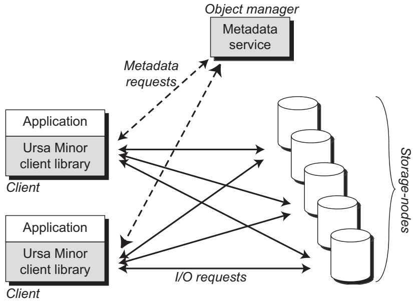
*Abd-El-Malek, Courtright, Cranor, Ganger, Hendricks, Klosterman, Mesnier, Prasad, Salmon, Sambasivan, Sinnamohideen, Strunk, Thereska, Wachs & Wylie*

Proceedings of the 4th USENIX Conference on File and Storage Technologies (FAST '05). San Francisco, CA. December 13-16, 2005.

No single encoding scheme or fault model is right for all data. A versatile storage system allows them to be matched to access patterns, reliability requirements, and cost goals on a per-data item basis. Ursa Minor is a cluster-based storage system that allows data-specific selection of, and on-line changes to, encoding schemes and fault models. Thus, different data types can share a scalable storage infrastructure and still enjoy specialized choices, rather than suffering from "one size fits all." Experiments with Ursa Minor show performance benefits of 2-3x when using specialized choices as opposed to a single, more general, configuration. Experiments also show that a single cluster

continued on page 17

continued from page 16



Ursa Minor high-level architecture. Clients use the storage system via the Ursa Minor client library. The metadata needed to access objects is retrieved from the object manager. Requests for data are then sent directly to storage-nodes.

supporting multiple workloads simultaneously is much more efficient when the choices are specialized for each distribution rather than forced to use a "one size fits all" configuration. When using the specialized distributions, aggregate cluster throughput increased by 130%.

## Empirical Analysis of Rate Limiting Mechanisms

*Wong, Bielski, Studer & Wang*

8th International Symposium on Recent Advances in Intrusion Detection (RAID 2005) September 7-9, 2005, Seattle, Washington.

One class of worm defense techniques that received attention of late is to "rate limit" outbound traffic to contain fast spreading worms. Several proposals of rate limiting techniques have appeared in the literature, each with a different take on the impetus behind rate limiting. This paper presents an empirical analysis on different rate limiting schemes using real traffic and attack traces from a sizable network. In the analysis we isolate and investigate the impact of the critical parameters for each scheme and seek to understand how these parameters might be set in realistic network settings. Analysis shows that using DNS-based rate limiting has substantially lower error rates than schemes based on other traffic statistics. The analysis additionally brings to light a number of issues with respect to rate limiting at large. We explore the impact of these issues in the context of general worm containment.

## A Study of Mass-mailing Worms

*Wong, Bielski, McCune & Wang*

WORM'04, October 29, 2004, Washington, DC, USA.

Mass-mailing worms have made a significant impact on the Internet. These worms consume valuable network resources and can also be used as a vehicle for DDoS attacks. In this paper, we analyze network traffic traces collected from a college campus and present an in-depth study on the effects of two mass-mailing worms, SoBig and MyDoom, on outgoing traffic. Rather than proposing a defense strategy, we focus on studying the fundamental behavior and characteristics of these worms. This analysis lends insight into the possibilities and challenges of automatically detecting, suppressing and stopping mass-mailing worm propagation in an enterprise network environment.

## Accelerating Database Operations Using a Network Processor

*Gold, Ailamaki, Huston & Falsafi*

ACM Int'l Workshop on Data Management on New Hardware (DaMoN), Baltimore, MD, June 12, 2005.

Database management systems (DBMSs) do not take full advantage of modern microarchitectural resources, such as wide-issue out-of-order processor pipelines. Increases in processor clock rate and instruction-level parallelism have left memory accesses as the dominant bottleneck in DBMS execution. Prior research indicates that simultaneous multithreading (SMT) can hide memory access latency from a single thread and improve throughput by increasing the number of outstanding memory accesses. Rather than expend chip area and power on out-of-order execution, as in current SMT processors, we demonstrate the effectiveness of using many simple processor cores, each with hardware support for multiple thread contexts. This paper shows an existing hardware architecture—the network processor—already fits the model for multi-threaded, multi-core execution. Using an Intel IXP2400 network processor, we evaluate the performance of three key database operations and demonstrate improvements of 1.9X to 2.5X when compared to a general purpose processor.

## A Computational Database System for Generating Unstructured Hexahedral Meshes with Billions of Elements

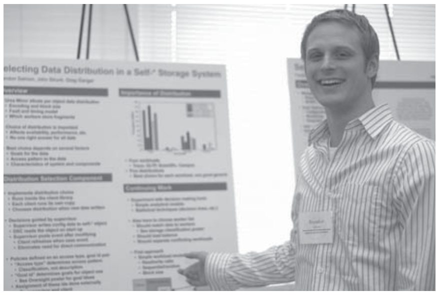
*Tu & O'Hallaron*

SC2004, November 6-12, 2004, Pittsburgh, PA USA

For a large class of physical simulations with relatively simple geometries, unstructured octree-based hexahedral meshes provide a good compromise between adaptivity and simplicity. However, generating unstructured hexahedral meshes with over 1 billion elements remains a challenging task. We propose a database approach to solve this problem. Instead of merely storing generated meshes into conventional databases, we have developed a new kind of software system called Computational Database System (CDS) to generate meshes directly on databases. Our basic idea is to extend existing database techniques to organize and index mesh data, and use database-aware algorithms to manipulate database structures and generate meshes. This paper presents the design, implemen-

continued on page 20

continued from page 15



Brandon Salmon discusses his work on "Data Distribution in Self-* Storage Systems" at the 2005 PDL Spring Industry Visit Day.

model on performance evaluation. In particular, we investigate the impact of choosing a closed versus an open system model, a question that has received little attention in the past.

## DISSERTATION ABSTRACT:

### A Read/Write Protocol Family for Versatile Storage Infrastructures

*Jay Wylie, ECE*

*Carnegie Mellon University Ph.D. Dissertation CMU-PDL-05-108, August 24, 2005.*

Today, high performance, high reliability storage systems are heavily engineered, very specialized, monolithic systems. The cost of such systems, in terms of raw storage capacity per dollar, is much more expensive than commodity storage. Storage bricks, scalable storage components built from commodity components, could reduce the cost of high performance, high reliability storage.

We have developed the Read/Write Protocol Family that enables a versatile storage infrastructure to be built out of storage bricks. Where versatile means "able to meet radically different fault-tolerance and performance requirements without modifying storage bricks". The Read/Write Protocol Family incorporates standard storage techniques (e.g., replication, striping, and RAID) with techniques for additional fault-tolerance (erasure codes and Byzantine fault-tolerance) and scalability (witnesses and quorum systems). Evaluation of a prototype storage system based on the Read/Write Protocol Family shows that versatility can be provided efficiently and that a single storage infrastructure can meet diverse fault-tolerance and performance requirements. As well, the prototype allows the real performance and cost trade-off of different failure assumptions to be explored.

## DISSERTATION ABSTRACT:

### Performance Modeling of Storage Devices Using Machine Learning

*Mengzhi Wang, CS*

*Carnegie Mellon University Ph.D. Dissertation, September 29, 2005.*

Performance models of system components are one of the key elements in building systems that can automatically allocate resources to deliver optimal performance. This thesis explores the feasibility of using machine learning techniques to build black-box models for storage devices. The models are constructed through "training", during which the model construction algorithm observes storage devices under a set of training traces and builds the models based on the observations. The main advantage of the approach is the automated model construction algorithm, in addition to the high efficiency in both computation and storage.

In our design, the models represent an I/O workload as vectors, and model its performance on storage devices as functions over the vectors using a regression tool. Therefore, it is important to design effective vector representations to capture important workload characteristics, in addition to choosing an accurate and efficient regression tool. Moreover, the quality of the training traces plays an important role in the final models, too, because the models base their predictions on the training data. We have proposed the entropy plot to characterize the spatio-temporal behavior of I/O workloads and the PQRS model to generate traces of given locality to augment existing work in workload characterization. This thesis evaluates the effectiveness of these techniques in implementing the learning-based device models.

The main conclusion is that with current progress in workload characterization, it is probably possible to build accurate and efficient models if one allows online model adaptation. When the training and testing traces are similar, the learning-based models can offer decent predictions. Offline training using synthetic traces, however, is less effective as synthetic trace generators need further improvements in generating high-quality training traces.

## DISSERTATION ABSTRACT:

### D-SPTF: Decentralized Request Distribution in Brick-based Storage Systems

*Christopher Lumb, ECE*

*Carnegie Mellon University Ph.D. Dissertation, November 19, 2004.*

Most current storage systems, including direct-attached disks, RAID arrays, and network filers, are centralized: they have a central point of control, with global knowledge of the system, for making data distribution and request scheduling decisions. This central control allows for good cache performance, load balancing and scheduling efficiency. However, many now envision building storage systems out of collections of federated bricks connected by high-performance networks. The goal of brick based storage is a system that has incremental scalability, parallel data transfer, and low cost. However with bricks there is no centralized point of control to provide request distribution. This lack of central control makes achieving good scheduling efficiency, load balancing and cache performance

continued on page 22

*continued from page 11*

types can share a scalable storage infrastructure and still enjoy specialized choices, rather than suffering from "one size fits all". This first system is being extended with detailed instrumentation and various automation agents to become the system (named Ursa Major) that we have designed for deployment. We hope to begin serving data for our first customer in the near future, and improve reliability and performance (and automation) iteratively thereafter.

### Early Contributors

The DCO is a huge undertaking, and even the initial steps have involved many. PDL has taken the lead in the design and early creation of the DCO, but many others at CMU and elsewhere continue to play crucial roles. CyLab, CMU's adminsitration, and CMU's central facilities organization (FMS) have all contributed knowledge and resources for the construction and power/cooling infrastructure needs. APC Corporation has partnered with us on the engineering plans and is also donating state-of-the-art power and cooling equipment. Several of PDL's sponsoring companies (Intel, IBM, and Seagate) and government sponsors (ARO, AFOSR, and DARPA) have provided equipment for early software development, especially of the self-* storage prototypes. And, of course, support for a lot of the research already comes from the PDL Consortium, NSF, ARO, AFOSR, and DARPA.

### Finale

The Data Center Observatory is a rare opportunity to gain firsthand, observation-based insight into the sources of operational costs and a real data center in which to deploy and evaluate new data center technologies. Things are off to a good start, but we're going to need a lot of help to successfully follow through on the DCO vision. We think it's absolutely worth doing, both for the technical knowledge that will be gained and the unprecedented student training that will be enabled. We look forward to working with industry partners and government sponsors in following this path and pushing the frontiers of data center automation and cost-effectiveness.

# NEW PDL FACES

## David Andersen

David Andersen is an assistant professor of computer science at Carnegie Mellon University. He completed his Ph.D. at MIT in December 2004, his dissertation discussing "Improving End-to-End Availability Using Overlay Networks." Prior to that, he received an MS in CS from MIT in 2001, and BS degrees in biology and computer science from the University of Utah. In 1995, he co-founded ArosNet, an Internet Service Provider in Salt Lake City, Utah. In his spare time, David enjoys rock climbing, cycling and running.

David's research at CMU focuses on networks and distributed systems, with an eye towards improving the availability and performance of Internet-based systems. His Highly Available Internet Architecture project examines methods of improving the availability and security of the Internet, without compromising the fundamental flexibility that underlies its success. Early work includes enhancing end-hosts' ability to select between paths through the Internet, and permitting hosts or networks greater control over the traffic they receive, in a way that is enforceable deep inside the network. This project emphasizes real-world measurements from a CMU-based Internet testbed as well as public testbeds to understand the problems facing today's network.

Collaborating with researchers at Intel, the Data-Oriented Transfer project is examining a new way to structure Internet applications that perform bulk transfers. Instead of performing the transfers themselves, these applications pass their data to a transfer service that performs the transfer on their behalf. The transfer service serves as a locus for the development and deployment of novel transfer techniques. Initial efforts include merging e-mail delivery with peer-to-peer techniques (e.g., collaborative delivery of large file attachments to multiple receivers), and developing transfer techniques to improve the performance of applications when the underlying network layers perform poorly.

The Opportunistic Resource Use in Wireless Networks project is investigating better ways to make use of wireless networks by explicitly exploiting concurrent multi-path transfers, by opportunistically caching overheard traffic, and by taking advantage of quiescent periods to preemptively transfer data across the network. This project is attempting to turn one of the weaknesses of wireless networks–a shared broadcast medium–into a strength, by taking advantage of the processing and storage power available at nodes to avoid expensive wireless communication.

*continued from page 17*

tation, and evaluation of a prototype CDS named Weaver, which has been used successfully by the CMU Quake project to generate queryable high-resolution finite element meshes for earthquake simulations with up to 1.22B elements and 1.37B nodes.

## The Seductive Appeal of Thin Clients

*Tolia, Andersen & Satyanarayanan*

Carnegie Mellon University School of Computer Science Technical Report CMU-CS-05-151, February, 2005.

Interest in thin clients is very high today because of frustration with the growing total cost of ownership of personal computers. Unfortunately, thin clients may not meet the usability goal of crisp interactive response. This paper shows that the adequacy of thin-client computing is highly variable, and depends on both the application and the available network quality. For intensely interactive applications, the tight control of end-to-end network latency required by thin clients may be hard to guarantee at large scale. The paper advocates the concept of stateless thick clients, and describes how they may reduce total cost of ownership while preserving good interactive performance.

Bianca Schroeder, now a post-doc with PDL, discusses "QoS for Databases" at the 2004 PDL Retreat.

## Correctness of the Read/Conditional-Write and Query/Update Protocols

*Abd-El-Malek, Ganger, Goodson, Reiter & Wylie*

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-05-107, September, 2005.

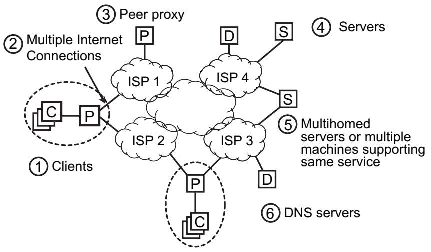
The Read/Conditional-Write (R/CW) protocol provides linearizable reads and conditional-writes of individual objects. A client's conditional-write of an object succeeds only if the object has not been conditionally-written since it was last read by the client. In this sense, R/CW semantics are similar to those of a compare-and-swap register. If a conditional-write does not succeed, it aborts. The R/CW protocol supports multi-object reads and conditional-writes; such operations are strictly serializable. A variant of the R/CW protocol, the Query/Update (Q/U) protocol, provides an operations-based interface to clients: clients invoke query and update methods on objects rather than reading and writing objects in their entirety. The R/CW and Q/U protocols are correct in the asynchronous timing model and tolerate Byzantine failures of clients and servers.

## Improving Web Availability for Clients with MONET

*Andersen, Balakrishnan, Kaashoek & Rao*

NSDI 2005, Boston, MA, May 2005.

Despite the increasing degree of multi-homing, path and data redundancy, and capacity available in the Internet, today's clients experience outage rates of a few percent when accessing Web sites. MONET ("Multi-homed Overlay NETwork), is a new system that improves client availability to Web sites using a combination of link multi- homing and a cooperative overlay network of peer proxies to obtain a diverse collection of paths between clients and Web sites. This approach creates many potential paths between clients and Web sites, requiring a scalable way to selecting a good path. MONET solves this problem using a waypoint selection algorithm, which picks a good small subset of all available paths to actively probe.


3 Peer proxy
2 Multiple Internet Connections
P
D
S
4 Servers
ISP 1
ISP 4
C
P
S
ISP 2
ISP 3
5 Multihomed servers or multiple machines supporting same service
1 Clients
P
D
6 DNS servers
C

The MONET environment. Clients (1) contact Web sites via a local MONET proxy. That local proxy may be multihomed with multiple local interfaces (2), and may also route requests through remote peer proxies (3). Clients wish to communicate with web sites (4), which may be themselves multihomed or spread over multiple machines (5). Web sites must be located using DNS (6); DNS servers are typically replicated over multiple machines.

MONET runs on FreeBSD, Linux, and Mac OS X, and is deployed at six different sites. These installations have been running MONET for over one year, serving about fifty users on a daily basis. Our analysis of proxy traces shows that the proxy network avoids between 60% and 94% of observed failures, including access link failures, Internet routing problems, persistent path congestion, and DNS failures. The proxy avoids nearly 100% of failures due to client and wide-area network failures, with negligible overhead.

*continued from page 9*

West Coast Campus provided a short history of storage innovation in Pittsburgh beginning with the Information Technology Center (ITC) in the early 80s, and the creation of the Andrew File System (AFS), which later became a product from Transarc and later IBM. AFS in turn inspired Coda, one basis for the distributed storage work now continuing at Intel Labs–Pittsburgh; led to Multi-Resident AFS (MR-AFS), still in use at the Pittsburgh Supercomputing Center (PSC); and inspired the global file system developed by Spinnaker Networks (now Network Appliance). The DSSC was founded in 1990 to study advanced magnetic recording, inventing some of the basic technology and training many of the technical innovators who are still increasing the capacity of disk drive storage. The DSSC provided early support for the Parallel Data Laboratory (PDL), which today leads the academic community in research into storage systems and originated the technology for Network Attached Secure Disks (NASD), a core component of the products currently being developed by Panasas. The DSSC also led to the founding of Seagate Research, which provides central R&D for the world's largest disk drive maker. The Pittsburgh Digital Greenhouse (PDG) supports commercialization and technology transfer, including storage technology, in the greater Pittsburgh region.

—with info from SC04 technical program notes

### October 2004
### CMU & PDL Host Posix Extensions Workshop

Carnegie Mellon University and the Parallel Data Lab hosted a Posix Extensions Workshop on November 8. The goal for the workshop was to achieve a well accepted by industry POSIX I/O API extension, or set of extensions, to make the POSIX I/O API more friendly to HPC, clustering, parallelism, and high concurrency applications. The meeting was held in conjunction with SuperComputing 2004 and covered the initial mechanics for how the POSIX API is to be extended, ideas for extensions, and the formation of a plan of attack, organization, and mapping of next steps.

### September 2004
### Spiros Papadimitriou Named a Siebel Scholar

Congratulations to Spiros Papadimitriou, who has been selected as a Siebel Scholar, providing him with one year of financial (tuition plus stipend) support. These scholarships are funded from an endowment set up by the Siebel Corporation.

## YEAR IN REVIEW

*continued from page 4*

**April 2005**

- ❖ Terrence Wong presented "Comparison-based File Server Verification" at USENIX05 in Anaheim, CA (NFS TEE project).

**March 2005**

- ❖ Christos Faloutsos spoke on "Data Mining Using Fractals and Power Laws" in the Wright State University, Dept. of CS Distinguished Lecturer Series. He also gave this lecture at Boston U. and Duke.

**February 2005**

- ❖ Craig Soules proposed his Ph.D. research "Automating Attribute Assignment Using Context to Assist in Personal File Retrieval."

**January 2005**

- ❖ Shuheng Zhou presented her paper "On Hierarchical routing in Doubling Metrics" at SODA 05 in Vancouver, BC, Canada.

**December 2004**

- ❖ Bill Courtright rejoins the PDL as Executive Director.
- ❖ Julio López, Niraj Tolia, Greg Ganger, Brandon Salmon, Eno Thereska, Mike Mesnier, Mike Abd-El-Malek, Craig Soules, James Hendricks, Jay Wylie attended OSDI 2004 in San Francisco.
- ❖ David Petrou successfully defended his dissertation on "Cluster Scheduling for Explicitly-Speculative Tasks."
- ❖ Greg spoke at Advanced Storage System Workshop organized by researchers from the Tokyo Institute of Technology and NAIST.

**November 2004**

- ❖ Chris Lumb defended his dissertation on "D-SPTF: Decentralized Request Distribution in Brick-based Storage Systems" and joined DataDomain.
- ❖ PDL participated in the "Pittsburgh – Epicenter of Storage" reception at SC'04 in Pittsburgh.
- ❖ POSIX extensions workshop held at CMU, attended by Garth Gibson, Greg Ganger, and several PDL grad students.

**October 2004**

- ❖ Christos Faloutsos presented "Advanced Data Mining Tools" as a Distinguished Lecturer at Univ. of Illinois, U-C.
- ❖ Greg spoke at the Intel Autonomic Summit in Portland.

**September 2004**

- ❖ 12th Annual PDL Retreat and Workshop.
- ❖ pNFS workshop held at Carnegie Mellon, attended by Garth Gibson, Greg Ganger, and several PDL graduate students.

continued from page 18

a challenge in decentralized brick based storage.

Distributed Shortest-Positioning Time First (D-SPTF) is a decentralized request distribution protocol designed to address this problem. D-SPTF exploits high-speed interconnects to dynamically select which server, among those with a replica, should service each read request. In doing so, it simultaneously balances load, exploits the aggregate cache capacity, and reduces positioning times for cache misses. For network latencies of up to 0.5ms, D-SPTF performs as well as would a hypothetical centralized system with the same collection of CPU, cache, and disk resources. Compared to a popular decentralized approach, hash-based request distribution, D-SPTF achieves up to 65% higher throughput and adapts more cleanly to heterogeneous server capabilities.

## DISSERTATION ABSTRACT:

### Staged Database Systems

*Stavros Harizopoulos, CS*

*Carnegie Mellon University Ph.D. Dissertation, September 19, 2005.*

Advances in computer architecture research yield increasingly powerful processors which can execute code at a much faster pace than they can access data in the memory hierarchy. Database management systems (DBMS), due to their intensive data processing nature, are in the front line of commercial applications which cannot harness the available computing power. To prevent processors from idling, a multitude of hardware mechanisms and software optimizations have been proposed. Their effectiveness, however, is limited by the sheer volume of data accessed and by the unpredictable sequence of memory requests.

PDL members stuffing binders in preparation for the 2004 PDL retreat and Workshop.

This Ph.D. dissertation introduces Staged Database Systems, a new software architecture for optimizing data and instruction locality at all levels of the memory hierarchy. The key idea is to break database request execution in stages and process a group of sub-requests at each stage. Group processing at each stage allows for a context-aware execution sequence of requests that promotes reusability of both instructions and data. The Staged Database System design requires only a small number of changes to the existing DBMS codebase and provides a new set of execution primitives that allow software to gain increased control over what data and instructions are accessed, when, and by which requests. The central thesis is the following: “By organizing and assigning system components into self-contained stages, database systems can exploit instruction and data commonality across concurrent requests thereby improving performance.”

## DISSERTATION ABSTRACT:

### Parameter-Free Spatial and Stream Mining

*Spiros Papadimitriou, CS*

*Carnegie Mellon University Ph.D. Dissertation, August, 2005.*

Data mining is the extraction of knowledge from large amounts of data and brings together the fields of databases, machine learning and statistics. The need to turn vast collections of data into useful knowledge has fueled the development of data mining techniques. From a database perspective, the emphasis is often placed on scalability and efficiency. Practical approaches can afford only a single or, at best, a few passes over the data, i.e., the algorithmic complexity must be linear with respect to dataset size.

Recently, in addition to the data warehouse model where data from multiple sources are integrated into a large store, the streaming model is emerging as an alternative data processing paradigm. Several applications produce a continuous stream of data (e.g., phone call detail records, web clickstreams or sensor measurements) that is too large to store in its entirety. Therefore, in stream mining we are allowed only a single pass over the data, without random access. Upon arrival of new observations, we have to incrementally update the data model. Furthermore, space complexity must be sublinear with respect to dataset size.

In this thesis we develop spatial and stream mining tools for discovery of interesting patterns. These patterns summarise the data, enable forecasting of future trends and spotting of anomalies or outliers. Beyond the emphasis on efficiency and scalability, we focus on simplifying or eliminating user intervention. We show that multi-resolution analysis (i.e., examining the data at multiple resolutions or scales) is a powerful tool towards these goals. In particular, for spatial data we employ the correlation integral. For time series streams we use the wavelet transform and related techniques. Furthermore, we leverage tools from signal processing (again wavelets and, also, subspace tracking algorithms) to extract patterns from streams. Finally, we also employ compression principles coupled with multi-level partitionings to automatically cluster spatial data.

The first two parts of this thesis focus on spatial mining methods. In the first part we examine homogeneous spatial data, where all points belong to one

continued on page 23

continued from page 22

class. In the second part we examine heterogeneous spatial data, where the points may belong to two or more different classes (e.g., species, galaxy types, etc). Finally, in the third part we focus on numerical, time series streams and mining techniques for both single and multiple streams.

### THESIS ABSTRACT:

### A Read/Write Protocol Family for Versatile Storage Infrastructures

*Mike Abd-El-Malek, ECE*
*Carnegie Mellon University Masters Thesis, August 24, 2005.*

Please refer to the abstract of the paper *Lazy Verification in Fault-Tolerant Distributed Storage Systems* under Recent Publications.

### THESIS ABSTRACT:

### Layout Characterization and Modeling for Modern Disk Drives

*John Bucy, ECE*

*Carnegie Mellon University Masters Thesis, August 9, 2005.*

Modern disk drives use increasingly complex schemes to map logical blocks (LBNs) to physical sectors. The variety and changes over time confound general disk characterization schemes that seek to deduce the mapping for any particular disk drive, whether for simulation model parameterization or per-request timing prediction. This technical report describes a general structure for concisely and accurately capturing logical-to-physical mappings.

Integrated into the DIXtrac disk characterization tool, it has been used successfully for over ten different disk makes/models. With the associated modeling software, integrated into DiskSim's diskmodel library, it maintains DiskSim's very high accuracy while enabling modern disks to be modeled.

### THESIS PROPOSAL:

### Automating Attribute Assignment Using Context to Assist in Personal File Retrieval

*Craig Soules, CS*

*School of Computer Science, Carnegie Mellon University, Feb. 25, 2005.*

Context information consists of both how a user perceives a piece of data, as well as how the user perceives the connection between data. A recent user study showed that most users use context information to locate their data during a search, however most organizational systems and search tools do not take this into account. In my work, I propose schemes to automatically gather context information from user activity and use that to generate file attributes for use in organization and search tools. By increasing the number of available attributes, I believe that such tools can be made more effective.

### THESIS PROPOSAL:

### An Instrumentation and Performance Querying Framework for Informed Tuning in a Self-managing System

*Eno Thereska, ECE*

*School of Computer Science, Carnegie Mellon University, July 6, 2005.*

In this thesis, I argue that systems should incorporate the ability to answer *What...if...* questions about their own performance. *What...if...* support converts complex tuning and policy decision into simpler search-based approaches. With them, iterating over a defined search space can produce a near-optimal answer. I believe, and plan to demonstrate, that a common framework, designed into the system from the beginning, can address a broad range of performance tuning problems in systems. Such a framework relies on instrumentation to account for per-workload, per-resource demand and *What...if...* modules that understand the system and are built-in from the start. Such accounting is not possible from outside the system, and hence, common approaches used in today's systems are limited in their effectiveness.

Erik Riedel of Seagate Technology in discussion with Stavros Harizopoulos, PDL Grad Student, and Jim Williams of Oracle at the 2004 PDL Retreat.

### THESIS PROPOSAL:

### Automating Attribute Assignment Using Context to Assist in Personal File Retrieval

*Niraj Tolia, ECE*

*Electrical & Computer Engineering, Carnegine Mellon University, June 3, 2005.*

The research will investigate the use of Content Addressable Storage (CAS) to address bulk data in conventional WAN-based client-server architectures. This thesis states that the use of CAS can be invaluable in breaking the dependency of clients on the server(s) as the sole data source. In particular, these techniques help in improving performance in bandwidth constrained environments. The thesis will show that CAS significantly improves performance without invasive architectural changes and while preserving the semantics of the original client-server system. Further, CAS provides a natural mechanism to enable the opportunistic use of non-networked data sources such as portable storage.

*Continued from page 13*

scheme for unstructured tetrahedral meshes, a common data organization for simulations that decomposes the continuous three-dimensional problem space into a collection of discrete pyramid-shaped elements. A tetrahedral mesh models a problem domain by decomposing it into tetrahedra or pyramid shaped elements and is well suited to modeling earthquake ground movement simulations. Figure 2(a) shows a part of a mechanical component mesh model, and Figure 2(b) illustrates a constituent tetrahedral element. The element endpoints, called the nodes, are the discrete points on which the simulation computes physical parameter values, such as earthquake ground velocities. Tetrahedral elements may have varying sizes, angles and orientations. When dealing with complex geometries that require variable resolutions, the tetrahedral mesh is a powerful modeling tool due to its unique flexibility in defining arbitrarily-shaped elements.

The most important query type for tetrahedral mesh datasets is the point query: given a point, the query returns the element containing the point, along with its corresponding nodes (the element's endpoints). Postprocessing and visualization applications use point queries in order to interpolate the value of a physical parameter on the particular query point, given the values computed by the simulation at the nodes.

This general functionality is vital to virtually every application that requires values at points that do not coincide with the input mesh node set. Point query performance is critical for applications like visualization, which require interactive rendering rates (less than 1s per frame). High frame rates are impossible to achieve on large-scale mesh datasets without efficient indexing techniques. Unfortunately, the pyramid-based geometry of tetrahedral meshes, while increasing their expressive power, makes developing effective indexing methods a challenging task.

The most popular and conceptually simple indexing technique is based on approximating each object (in this case each pyramid-shaped element) by its Minimum Bounding Rectangle (MBR). The MBRs are then indexed using an R-Tree. This offers sub-optimal performance when applied to tetrahedral meshes because the MBR does not efficiently capture pyramid-shaped elements. Other multidimensional indexing techniques, involving clipping or space-filling curves have similar performance problems.

To counter the performance problems of existing database indexing techniques a new query processing algorithm for point queries on large-scale tetrahedral meshes has been developed, called Directed Local Search (DLS), to provide better performance than existing multidimensional indexing techniques, independent of the mesh geometry. It has been demonstrated that DLS has superior performance compared to existing techniques, reducing the number of page accesses by up to an order of magnitude. In addition, DLS can be easily integrated in existing relational DBMS, without requiring new and exotic access methods. In contrast to many sophisticated multidimensional indexing techniques, DLS also requires minimal preprocessing.

The DLS implementation has been evaluated on five real mesh datasets, one of which is used for earthquake modeling. The experiment demonstrated that DLS outperforms R-Tree based indexing for point location queries, by nearly eliminating R-Tree page accesses and achieving similar performance for elements and nodes accesses.

## CONTRIBUTORS

Anastassia Ailamaki, Greg Ganger, Gerd Heber (Cornell Theory Center), Julio López, Dave O'Hallaron, Stratos Papadomanolakis and Tianka Tu.

## REFERENCES

Tu, T. and O'Hallaron, D. A Computational Database System for Generating Unstructured Hexahedral Meshes with Billions of Elements. SC2004, Nov. 6-12, 2004, Pittsburgh, PA.

Papadomanolakis, et al. Efficient Query Processing for Unstructured Tetrahedral Meshes. In submission.

PDL Retreat 2004 attendee group photo.