THE
# PDL Packet
THE NEWSLETTER ON PARALLEL DATA SYSTEMS • FALL 2000
http://www.pdl.cs.cmu.edu

AN INFORMAL PUBLICATION FROM A UNIVERSITY RESEARCH COMMUNITY DEVOTED TO ADVANCING THE STATE OF THE ART IN STORAGE SYSTEMS AND TO EFFICIENTLY INTEGRATING STORAGE INTO PARALLEL AND DISTRIBUTED FILE SYSTEMS, HIGH BANDWIDTH NETWORKS AND COMPUTER CLUSTERS.

## CONTENTS



## CONSORTIUM MEMBERS

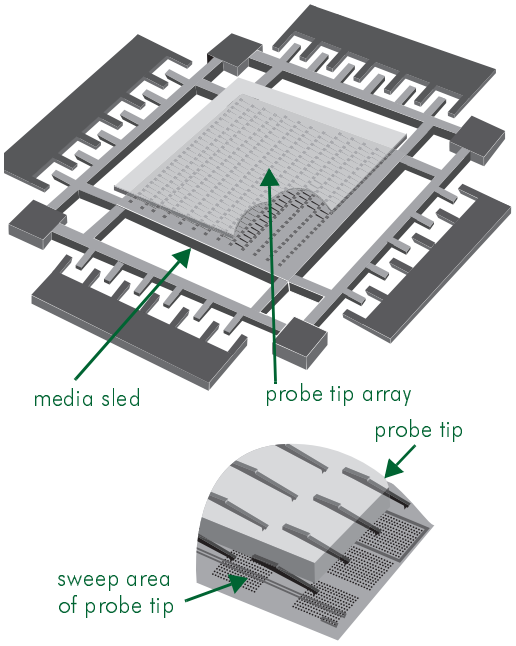
EMC Corporation
Hewlett Packard Labs
Hitachi, Ltd.
IBM
Infineon Technologies
Intel Corporation
LSI Logic
Lucent Technologies
MTI Technology Corp.
Novell, Inc.
Panasas, Inc.
Platys Communications
Procom Technology
Quantum Corporation
Seagate Technology
Sun Microsystems
Veritas Software Corp.
3Com Corporation

# MicroElectroMechanical Systems (MEMS)-Based Storage

http://www.lcs.ece.cmu.edu/research/MEMS/
http://www.chips.ece.cmu.edu/

*David Nagle*

Imagine a world where gigabytes of storage, 1,000s of MIPS, and gigabit/second networking are merged into a single chip smaller than a quarter. This is the vision of CMU's new Center for Highly Integrated Information Processing and Storage Systems (CHIPS). CHIPS's goal is to revolutionize systems, creating low-cost embedded computers with multiple gigabytes of IC-based mass storage that will become a ubiquitous part of our everyday environment. IC-based mass storage devices will also enhance the security and archivability of data storage systems by enabling a tightly integrated coupling of storage and processing. Desktop and laptop computers architectures will evolve to incorporate IC-based mass storage into their memory hierarchies and exploit order-of-magnitude access times reduction, resulting in significant performance improvements. Even the capabilities of massively parallel computers will be enhanced under this vision as the small size of mass storage brings it closer to the processor, enabling dramatic performance improvements on applications ranging from data mining to fast FFTs.




Prototype MEMS-based data storage system.

The key to a true system-on-a-chip is the integration of gigabytes of non-volatile memory on a chip. To solve this problem, CMU researchers have turned to hybrid approaches that leverage the best of semiconductor memories and disk drives (IBM and HP have adopted similar approaches). From semiconductor memories, we adopt the wafer fabrication process to minimize unit costs. From disk drives, we adopt recording heads that use mechanical position to address data stored in a thin film material. For compatibility with silicon fabri-

... continued on pg. 12

# PDL 2000

## From the Director's Chair

DAVID NAGLE

Howdy from the great state of Pennsylvania. 2000 has been a very exciting year from the Parallel Data Lab. Greg Ganger has become co-director of the PDL, recognizing Greg's very significant and often un-noticed contributions. Todd Mowry, Garth Gibson, and Hui Zhang received National Science Foundation Information Technology and Research Grants. Hui also won the Sloan Foundation Fellowship. Khalil Amiri and Ian Stoica successfully defended their dissertations. Khalil is off to IBM Watson and Ion has joined the faculty at UC Berkeley. Craig Soules, one of PDL's new graduate students, received a USENIX fellowship and Greg and Jenny Ganger had a bouncing baby boy, nicknamed "The Rock." You'll find more detailed announcements on page 8.

2000 has brought numerous staff and student changes. After five years with the PDL, Patty Mackiewicz has joined Carnegie Technology, a subsidiary of Carnegie Mellon University that offers Internet-based learning and certification for software developers. Many thanks to Patty for all of her work for PDL and congrats on the new job. Jennifer Landefeld also left the PDL, moving on to join Panasas. Jeff Butler, Charles Hardin, Ed Hogan, Chris Sabol, and Mike Scheinholtz have all graduated and moved into industry. There are also lots of new faces. Karen Lindenfelser has accepted the job of PDL Business Manager, Stephen Miller has joined us as our lab secretary, and five new students have joined the lab. As well, Srinivasan Seshan, one of CMU's new faculty, is joining us at this retreat and will be presenting his work on mobile networking.

PDL continues to work on numerous research projects that span the storage spectrum from low-level storage device characterization to file systems and storage networking. Allow me to highlight some of the major research advanced during the last 12 months.

The MEMS-based Storage Systems project is finding a home in CMU's new Center for Highly Integrated Information Processing and Storage Systems (CHIPS). Lead by Rick Carley, CHIPS has a unique vision – a future in which computer systems have been revolutionized by the incorporation of non-volatile rewritable mass storage devices manufactured using parallel photo-lithographic integrated-circuit (IC) compatible fabrication techniques. The CHIPS vision predicts that computer systems will embrace non-volatile rewritable IC-based mass storage (IMS) devices because manufacturing mass storage devices entirely using IC-compatible fabrication methods will result in a quantum decrease in their entry cost, access time, volume, mass, power needs, failure rate, and shock sensitivity offering dramatically improved performance and the capability for many new applications.

Currently, CHIPS is working on building prototypes of these devices. In October, the CHIPS group sent out for fabrication a new media positioning system. Also, Mike Lu and David Guillou have built a prototype capable of precise z-height sensing and control. On the systems front, John Griffin and Steve

*... continued on pg. 3*

**THE PDL PACKET**

The Parallel Data Laboratory
School of Computer Science
Deptartment of Electrical &
Computer Engineering
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
VOICE 412•268•6716
FAX 412•268•3010
FAX ALTERNATE 412•268•5576

PUBLISHER
David Nagle

EDITOR
Joan Digney

The PDL Packet is published once per year and mailed to members of the Parallel Data Consortium. Copies are given to other researchers in industry and academia as well. 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.

### MISSION STATEMENT

To advance the state of the art in storage systems and integration into parallel and distributed file systems, high bandwidth networks, and computer clusters.

### CONTACTING US

#### WEB PAGES

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

Please see our web pages at http://www.pdl.cs.cmu.edu/PEOPLE for further contact information.

#### FACULTY

David Nagle (director)
412•268•3898
david.nagle@cs.cmu.edu

Greg Ganger (co-director)
greg.ganger@ece.cmu.edu

Christos Faloutsos
christos.faloutsos@cs.cmu.edu

Garth Gibson
garth.gibson@cs.cmu.edu

Seth Goldstein
seth.goldstein@cs.cmu.edu

Mor Harchol-Balter
mor.harchol-balter@cs.cmu.edu

Todd Mowry
todd.mowry@cs.cmu.edu

Hui Zhang
hui.zhang@cs.cmu.edu

#### STAFF MEMBERS

Karen Lindenfelser (pdl business administrator)
412•268•6716
karen@ece.cmu.edu
Mike Bigrigg
Shelby Davis
Joan Digney
Nat Lanza
Paul Mazaitis
Stephen Miller
Semih Oguz
Joe Ordia
Nitin Parab
Melissa Puryear

#### GRADUATE STUDENTS

Mehmet Bakkaloglu
Angela Demke Brown
Fay Chang
Shimin Chen
David Friedman
Garth Goodson
John Griffin
Stavros Harizopoulos
Andrew Klosterman
Lesley Leposo
Chris Lumb
Vijay Pandurangan
David Petrou
Jiri Schindler
Steve Schlosser
Craig Soules
John Strunk
Mengzhi Wang
Ted Wong
Jay Wylie
Shuheng Zhou

# PDL 2000, continued

*… continued from pg. 2*

Schlosser built a MEMS-based storage device simulator that can be customized to model a wide range of possible MEMS-based storage device configurations. You'll find more information on CHIPS and the MEMS research on the PDL's "Members Only" web pages and at www.chips.ece.cmu.edu.

Todd Mowry and Garth Gibson are continuing their prefetching and cache management research with a National Science Foundation Information Technology and Research grant to fund supporting "Static and Dynamic Techniques for Hiding Latency in Data-Intensive Application." The work seeks to hide I/O latency by developing automatic methods that extract program-specific knowledge of I/O access patterns and provide the runtime support necessary to fully exploit this knowledge. Central to the work is a merging of compilation and program transformation technology with operating system resource management to better manage caching and prefetching algorithms and system resources. You'll find more information on this project in Fay Chang's 1999 SOSP paper, Angela Demke Brown's 1996 OSDI paper and Fay's upcoming dissertation, all available (or soon-to-be available) on the PDL web pages.

Greg Ganger's group has just released DIXtrac, a tool for automated disk drive characterization. Without human intervention, DIXtrac extracts over 100 disk parameters pertaining to disk geometry, data layout, disk mechanics, cache behavior, and various command processing overheads. DIXtrac characterizes either SCSI or FibreChannel disk drives by issuing SCSI commands directly to a disk and exercises specific test vectors consisting of read and write requests. DIXtrac has characterized 11 different disk models which are used to configure our DiskSim disk simulator. In comparison tests, the performance of the real disk and the simulated disk varied by less than 5%. DIXtrac runs as a user level process on a Pentium-based PC under Linux and is available to the PDC on our "Members Only" web page.

Finally, our storage networking research is focusing on how the network can best support SCSI-over-IP. We have built a working SCSI-over-IP prototype that ships all SCSI traffic over a TCP/IP or UDP/IP network, using Linux's SCSI Direct (SD) option to communicate with a SCSI disk drive. Our current prototype runs in both Linux and NetBSD, supports multiple initiators, and is currently being extended to support striping of data across multiple TCP/IP connections. The code is available from the PDL "Members Only" web page.

In addition to our research efforts, we continue to work with industry organizations. Garth is serving on the Storage Network Industry Association (SNIA) technical council, which is currently developing a storage architecture document (see www.snia.org for more information). Shortly after beginning our SCSI-over-IP work, we presented an overview of issues important to SCSI-over-IP to a pre-IETF working group meeting. This and many subsequent meetings between numerous industry people have led to an IETF (Internet Engineering Task Force) working group tasked to define the SCSI-over-IP standard (www.ece.cmu.edu/~ips has more information). Finally, our Network-attached Secure Disk (NASD) project has been highlighted in several books including "NFS Illustrated," by Brent Callagan, "Building Storage Networks," by Marc Farley, and "Parallel I/O for High-Performance Computing," by John May.

In writing this summary, I am always amazed at the research accomplishments PDL's faculty, students and staff achieve each year. As always, much of the credit goes to the staff and students who do the lion's share of the work. Congrats to the everyone on a very successful year.

## YEAR IN REVIEW

**November 2000**
- Steve Schlosser on "Designing Computer Systems with MEMS-based Storage." ASPLOS-IX, Cambridge, MA.

**October 2000**
- John Strunk on "Self-Securing Storage: Protecting Data in Compromised Systems," OSDI, San Diego, CA.
- John Griffin on "Operating System Management of MEMS-based Storage Devices," OSDI, San Diego, CA.
- Greg Ganger on "Towards Higher Disk Head Utilization: Extracting 'Free' Bandwidth From Busy Disk Drives," OSDI, San Diego, CA.

**September 2000**
- Khalil Amiri defends Ph.D. Dissertation: "Scalable and Manageable Storage Systems."
- Ion Stoica defends Ph.D. Dissertation: "A Stateless Core Approach for Scalable Internet Services."
- David Petrou speaks on "Easing the Management of Data-parallel Systems" at ACM SIGOPS European Workshop in Kolding, Denmark.
- Christos Faloutsos gives the Keynote address at COMLEX 2000, Univ. of Patras, Greece, on "Multimedia Indexing."
- Garth Gibson attends SNIA Technical Council meetings (in January, February, March and September).

**June 2000**
- Steve Schlosser on "Modeling and Performance of MEMS-Based Storage Devices." ACM SIGMETRICS 2000, Santa Clara, CA.

**May 2000**
- Christos Faloutsos gives invited talk on "Searching, Data Mining and Visualization of Multimedia Data" at Visual Databases VDB5, Fukuoka, Japan.
- Erik Riedel on "Data Mining on an OLTP System (Nearly) for Free," SIGMOD, Dallas, TX.
- PDS Spring Open House.

**April 2000**
- Khalil Amiri on "Highly Concurrent Shared Storage." ICDCS, Taipei, Taiwan.
- Mor Harchol-Balter on "Task Assignment with Unknown Duration" ICDCS, Taipei, Taiwan.
- Garth Gibson attends SNIA Technical Council meetings.

**March 2000**
- Garth Gibson participates in the Presidential Information Technical Advisory Council meeting on Open Source software development for high performance computing.

**February 2000**
- David Nagle speaks on Storage over IP at pre-IETF Working Group on Storage over IP. Santa Clara, CA.

**December 1999**
- David Nagle gives invited talk at MIT on NASD.
- Garth Gibson speaks on "NASD: Network-Attached Secure Disks," at the 1999 Seagate Research Conclave, Bloomington, MN.

**November 1999**
- PDS Retreat & Workshop.
- Final NSIC/NASD public meeting; first SNIA/OBSD meeting, Santa Clara, CA., Garth Gibson speaks on "Object-Based Storage Devices: Scalability and Aggregation," and "Object-Based Storage Devices: (Cryptographic) Security."
- Jim Gray of Microsoft; distinguished lecturer, CMU.

## NEW CONSORTIUM MEMBERS

Since last year's retreat, several new companies have joined (or rejoined) the Parallel Data Consortium:

- IBM (www.ibm.com)
- Lucent Technologies (www.lucent.com)
- Platys Communications (www.platys.com)
- Panasas, Inc. (www.panasas.com)
- Sun Microsystems (www.sun.com)
- Veritas Software Corporation (www. veritas.com)

We are looking forward to a seeing all of you at the 8th Annual Parallel Data Systems Workshop and Retreat at the Nemacolin Woodlands Resort this fall and to the beneficial exchange of ideas that occurs there between the PDL all the PDC members.

Opportunity to exchange ideas with industry guests on the Retreat Hike.

Ph.D. grad student, Khalil Amiri, busy keeping the industry sponsor happy!

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

*Lumb, Schindler, Ganger, Riedel & Nagle*

Appears in Proceedings of the 4th Symposium on Operating Systems Design and Implementation, San Diego, CA. October, 2000.

ABSTRACT

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

## Secure Continuous Biometric-Enhanced Authentication

*Klosterman & Ganger*

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

ABSTRACT

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

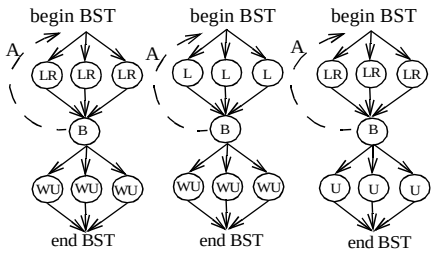
## Highly Concurrent Shared Storage

*Amiri, Gibson & Golding*

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

ABSTRACT

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




BST implementations with device-served locking and piggy-backing optimization. Node A represents a message exchange with a device. An L node denotes a lock operation; a U node stands for an unlock operation. LR represents the lock-and-devread operation, and WU represents the devwrite-and-unlock. The edges represent control dependencies. A B node represents a commit point at the host, where the host blocks until all preceding operations complete, restarting from the beginning if any of them fail.

## Data Mining on an OLTP System (Nearly) for Free

*Riedel, Faloutsos, Ganger & Nagle*

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

ABSTRACT

This paper proposes a scheme for scheduling disk requests that takes advantage of the ability of high-level functions to operate directly at individual disk drives. We show that such a scheme makes it possible to support a Data Mining workload on an OLTP system almost for free: there is only a small impact on the

*... continued on pg. 6*

*... continued from pg. 6*

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

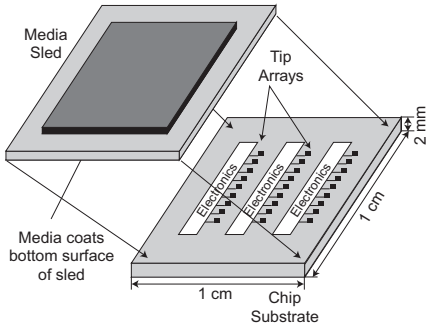
## Operating System Management of MEMS-Based Storage Devices

*Griffin, Schlosser, Ganger & Nagle*

Appears in Proceedings of the 4th Symposium on Operating Systems Design and Implementation, San Diego, CA, October 2000.

ABSTRACT

MEMS-based storage devices promise significant performance, reliability, and power improvements relative to disk drives. This paper compares and contrasts these two storage technologies and explores how the physical characteristics of MEMS-based storage devices change four aspects of operating system (OS) management: request scheduling, data placement, failure management, and power conservation. Straightforward adaptations of existing disk request scheduling algorithms are found to be appropriate for MEMS-based storage devices. A new bipartite data placement scheme is shown to better match these devices' novel mechanical positioning characteristics. With aggressive internal redundancy, MEMS-based storage devices can mask and tolerate failure modes that halt operation or cause data loss for disks. In addition, MEMS-based storage devices simplify power management because the devices can be stopped and started rapidly.




Components of a MEMS-based storage device. The media sled is suspended above an array of probe tips. The sled moves small distances along the X and Y axes, allowing the stationary tips to address the media.

## Automated Disk Drive Characterization

*Schindler & Ganger*

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

ABSTRACT

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

## Designing Computer Systems with MEMS-Based Storage

*Schlosser, Griffin, Nagle & Ganger*

Appears in Proceedings of the 9th International Conference on Architectural Support for Programming Languages and Operating Systems, 2000.

ABSTRACT

For decades the RAM-to-disk memory hierarchy gap has plagued computer architects. An exciting new storage technology based on microelectromechanical systems (MEMS) is poised to fill a large portion of this performance gap, significantly reduce system power consumption, and enable many new applications. This paper explores the system-level implications of integrating MEMS-based storage into the memory hierarchy. Results show that standalone MEMS-based storage reduces I/O stall times by 4-74X over disks and improves overall application runtimes by 1.9-4.4X. When used as on-board caches for disks, MEMS-based storage improves I/O response time by up to 3.5X. Further, the en-

*... continued on pg. 7*

*… continued from pg. 6*

ergy consumption of MEMS-based storage is 10-54X less than that of state-of-the-art low-power disk drives. The combination of the high-level physical characteristics of MEMS-based storage (small footprints, high shock tolerance) and the ability to directly integrate MEMS-based storage with processing leads to such new applications as portable gigabit storage systems and ubiquitous active storage nodes.

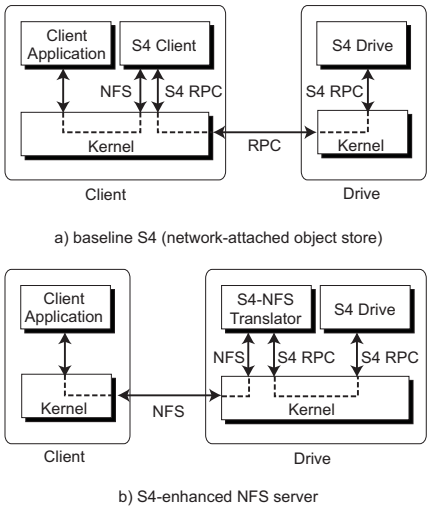
## Self-Securing Storage: Protecting Data in Compromised Systems

*Strunk, Goodson, Scheinholtz, Soules & Ganger*

Appears in Proceedings of the 4th Symposium on Operating Systems Design and Implementation. San Diego, CA. October, 2000.

ABSTRACT

Self-securing storage prevents intruders from undetectably tampering with or permanently deleting stored data. To accomplish this, self-securing storage devices internally audit all requests and keep old versions of data for a window of time, regardless of the commands received from potentially compromised host operating systems. Within the window, system administrators have this valuable information for intrusion diagnosis and recovery. Our implementation, called S4, combines log-structuring with journal-based metadata to minimize the performance costs of comprehensive versioning. Experiments show that self-securing storage devices can deliver performance that is comparable with conventional storage systems. In addition, analyses indicate that several weeks worth of all versions can reasonably be kept on state-of-the-art disks, especially when differencing and compression technologies are employed.




Two S4 configurations that provide self-securing storage via a NFS interface. (a) shows S4 as a network-attached object store with the S4 client daemon translating NFS requests to S4-specific RPCs. (b) shows a self-securing NFS server created by combining the NFS-to-S4 translation and the S4 drive.

## Active Disk Architecture for Databases

*Riedel, Faloutsos & Nagle*

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

ABSTRACT

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

*… continued on pg. 8*

Affordable grad student housing.

*… continued from pg. 7*

## Evaluation of Task Assignment Policies for Supercomputing Servers: The Case for Load Unbalancing and Fairness

*Schroeder & Harchol-Balter*

9th IEEE Symposium on High Performance Distributed Computing (HPDC '00), Pittsburgh, Pennsylvania, August 2000.

ABSTRACT

While the MPP is still the most common architecture in supercomputer centers today, a simpler and cheaper machine configuration is growing increasingly common. This alternative setup may be described simply as a "collection of multiprocessors" or a "distributed server system." This collection of multiprocessors is fed by a single common stream of jobs, where each job is dispatched to exactly one of the multiprocessor machines for processing.

The biggest question which arises in such distributed server systems is what is a good rule for assigning jobs to host machines: i.e. what is a good "task assignment policy." Many task assignment policies have been proposed, but not systematically evaluated under supercomputing workloads.

In this paper we start by comparing existing task assignment policies using a trace-driven simulation under supercomputing workloads. We validate our experiments by providing analytical proofs of the performance of each of these policies. These proofs also help provide much intuition. We find that while the performance of supercomputing servers varies widely with the task assignment policy, none of the above task assignment policies perform as well as we would like.

We observe that all policies proposed thus far aim to balance load among the hosts. We propose a policy which purposely unbalances load among the hosts, yet, counter intuitively, is also fair in that it achieves the same expected slowdown for all jobs – thus no jobs are biased against. We evaluate this policy again using both trace-driven simulation and analysis. We find that the performance of the load unbalancing policy is significantly better than the best of those policies which balance load.

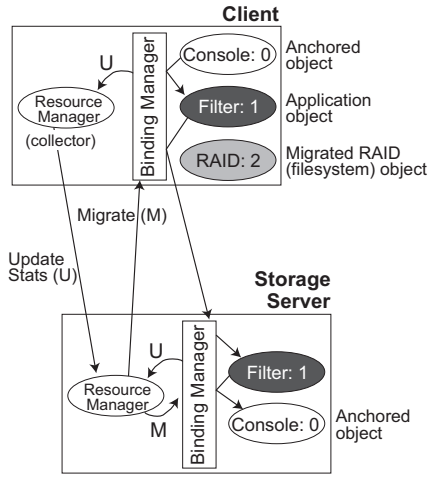
## Dynamic Function Placement for Data-Intensive Cluster Computing

*Amiri, Petrou, Ganger & Gibson*

Proceedings of the USENIX Annual Technical Conference, San Diego, CA, June 2000.

ABSTRACT

Optimally partitioning application and filesystem functionality within a cluster of clients and servers is a difficult problem due to dynamic variations in application behavior, resource availability, and workload mixes. This paper presents ABACUS, a run-time system that monitors and dynamically changes function placement for applications that manipulate large data sets. Examples of data-intensive workloads show the importance of proper function placement and its dependence on dynamic run-time characteristics, with performance differences frequently reaching 2-10X. We evaluate how well the ABACUS prototype adapts to run-time system behavior, including both long-term variation (e.g., filter selectivity) and short-term variation (e.g., multi-phase applications and inter-application resource contention). Experiments with ABACUS indicate that adaptation is possible in all of these situations and that it converges most quickly in those cases where the performance impact is most significant.




An ABACUS object graph, the principal ABACUS components, and their interactions. This figure shows a filter application accessing a striped file. Functionality is partitioned into objects. Dark ovals depict mobile objects, while clear ovals mark anchored objects. Inter-object method invocations are transparently redirected by the location transparent invocation component of the ABACUS run-time. This component also updates a local resource monitoring component on each procedure call and return from a mobile object (machine-local arrows labeled U). Clients periodically send digests of this collected information to the server. Resource managers at the server collect the relevant statistics and initiate migration decisions (arrows labeled M).

*… continued on pg. 9*

No wonder John gets so much work done – there are two of him!

*... continued from pg. 8*

## Easing the Management of Data-parallel Systems via Adaptation

*Petrou, Amiri, Ganger & Gibson*

Proceedings of the 2000 European SIGOPS Workshop, Kolding, Denmark, September 2000.

ABSTRACT

In recent years we have seen an enormous growth in the size and prevalence of data-mining workloads. We argue that high availability and fast turnaround for these workloads can only be realized by dynamically tuning a number of system parameters. Further, we argue that this tuning should be provided automatically by the system. We contribute a framework that enables the expression of a variety of data-parallel applications, but which is also sufficiently restricted so that the system can tune itself. This framework is part of the Abacus migration system, whose function placement algorithms are extended to reason about how many nodes should participate in a data-parallel computation, how to split up application objects among a client and server cluster, how often program state should be checkpointed, and the interaction (sometimes conflicting) between these questions.

## Modeling and Performance of MEMS-Based Storage Devices

*Griffin, Schlosser, Ganger & Nagle*

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

ABSTRACT

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

## Task Assignment with Unknown Duration

*Harchol-Balter*

20th International Conference on Distributed Computing Systems (ICDCS '00), Taipei, Taiwan, April 2000.

ABSTRACT

We consider a distributed server system and ask which policy should be used for assigning tasks to hosts. In our server, tasks are not preemptible. Also, the task's service demand is not known a priori. We are particularly concerned with the case where the workload is heavy-tailed, as is characteristic of many empirically measured computer workloads. We analyze several natural task assignment policies and propose a new one TAGS (Task Assignment based on Guessing Size). The TAGS algorithm is counterintuitive in many respects, including load unbalancing, non-work-conserving, and fairness. We find that under heavy-tailed workloads, TAGS can outperform all task assignment policies known to us by several orders of magnitude with respect to mean response time and mean slowdown, provided the system load is not too high. We also introduce a new practical performance metric for distributed servers called server expansion. Under the server expansion metric, TAGS significantly outperforms all other task assignment policies, regardless of system load.

## MEMS-Based Integrated Circuit Mass Storage Systems

*Carley, Ganger & Nagle*

Communications of the ACM, Vol. 43, No. 11 November, 2000. Special issue on Ultra-High-Density Data Storage.

ABSTRACT

At over 100% per year, the growth rate of the storage capacity of a disk drive continues to outpace the 60% per year growth rate of semiconductor memories, allowing disk drives to maintain and reinforce their position as the most cost effective non-volatile storage solution available. Unfortunately, decreases in disk drive access times have been minimal, creating a significant performance problem for common small-access workloads such as transaction workloads. And, the minimum entry cost of disk drives, though it has declined in recent years due to the decrease in the number of disks and heads in a disk drive, is still much too high for many consumer applications.

In order to pierce both the access time and entry cost barriers posed by disk drives, researchers have turned

*... continued on pg. 16*

**October 1, 2000**

### SCSI Object Based Storage Device Commands Set, Revision 3 Released

On October 1, the T10 working draft of the SCSI Object Based Storage Device Commands Set, Revision 3 was released. The SCSI command set is designed to provide efficient peer-to-peer operation of input/output logical units by an operating system using Object Based Storage commands. Objects designate entities in which computer systems store data. The purpose of the OSD abstraction is to assign to the storage device the responsibility for managing where data is located on the device. A downloadable pdf version of this document is available from the PDL Publications page.

**September 13, 2000**

### Two PDL Faculty Awarded NSF Information Technology Research (ITR) Grants

In September, the National Science Foundation announced the first awards under its new $90 million Information Technology Research (ITR) initiative. From the 1400 proposals the NSF received, 15% were selected to receive awards. The NSF has requested additional funding of $190 million for its fiscal 2001 ITR budget and the second NSF ITR competition has already begun.

Six research projects from Carnegie Mellon University were among the successful proposals and are slated to receive $5,513,665 over the next three to five years. Carnegie Mellon University faculty and research scientists had more proposals selected than any other academic institution, with two of the proposals coming from PDL Faculty.

Todd Mowry

Hui Zhang

Hui Zhang's research will investigate "Collaborative Research: Scalable Services for the Global Network." Todd Mowry will look at "Static and Dynamic Techniques for Latency Hiding in Data-Intensive Applications."

**September, 2000**

### NASD in the Classroom

Network Attached Secure Disks (NASD) is continuing to gain recognition and was overviewed in two textbooks this year. *NFS Illustrated* by Brent Callaghan, published by Addison-Wesley in 1999, *Building Storage Networks* by Marc Farley, published by McGraw Hill in 1999 and *Parallel I/O for High-Performance Computing* by John May, published by Morgan Kaufman all devoted several pages to the subject.

**Summer 2000**

### David Nagle Named Associate Director of CHIPS

Congratulations to Dave on his appointment to the position of Associate Director of the Center for Highly Integrated Information and Storage Systems (CHIPS). CHIPS is an interdisciplinary, systems-oriented research center created within the Department of Electrical and Computer Engineering at CMU that focuses on the creation and application of low-cost, highly integrated computing and mass data storage systems.

**September, 2000**

### PDL and CHIPS Lead in Special Issue of CACM

Papers detailing our research on MEMS-based Storage and Network Attached Storage are featured in the November issue of the Communications of the Association for Computing Machinery (CACM). The November issue is a special issue devoted to storage.

**August, 2000**

### New CS Faculty

We'd like to take this opportunity to welcome Srinivasan Seshan, one of our new CS faculty members. Srini received his Ph.D. from the University of California at Berkeley, California in 1995 and until recently was employed in the Networking and Security Department at IBM's Thomas J. Watson Research Center. His research interests are in network software for computer systems and he is currently working on new network protocols and services to support ubiquitous computing applications and wide-area distributed network applications.

**Summer 2000**

### Craig Soules Receives USENIX Scholarship

Our congratulations to Craig Soules, for being named a "USENIX Scholar" by the USENIX Association. As a part of this award, USENIX is supporting Craig's tuition and stipend this year. Craig is currently working as a CS graduate student on Self-Securing Storage Devices.

**Summer 2000**

### Hui Zhang on Leave

Hui has recently taken 2 years' leave from his position as Assistant Professor at CMU to serve as CTO at Turin Networks (www.turinnetworks.com) in Petaluma, CA. The company's products address the bandwidth scalability, multiservice performance and quality of service issues associated with the transformation of the public network to a multi-service broadband Internet.

*... continued on pg. 11*

*... continued from pg. 10*

**Summer 2000**

### Rick Carley at Startup

Professor Rick Carley is CTO and one of the co-founders of IC Mechanics, a new startup company that plans to develop and market very low cost inertial sensors for use in hard disk drives. IC Mechanics plans to sell rotational vibration sensors, active damping sensors, and inertial self servo writing sensors. Rick's email address is rick.carley@icmechanics.com.

John Strunk and Greg Ganger at Spring 2000 Convocation

**May 2000**

### ECE Students Graduate

Garth Goodson, John Griffin, Andy Klosterman, Steve Schlosser, Chris Lumb, John Strunk and Shuheng Zhou all received their M.S. degrees in ECE in May 2000 and are all continuing on for their Ph.D.s at CMU. Mike Scheinholtz received his M.S. in ECE and now works at Mirapoint. Chris Sabol received his M.S. and now works at 2Wire along with Charles Hardin who graduated this spring. Jeff Butler also received an M.S. in ECE this spring and now works at Panasas.

**March 25, 2000**

### Timothy Ganger (a.k.a. The Rock) Arrives

No question about it – Timothy is a Ganger. In true Ganger male fashion, he barely noted the passing of the official deadline, asked to know the final extended deadline (Sunday's induction), and slipped in just under it (8:57p.m. on Saturday, March 25).

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

Timothy has his mother's face, his father's ability to let people know when his drawers are messy, and Shaq-like size. Coming into the world at 22 inches and 9 pounds, 12 ounces, he's not going to let the other babies on the block push him around. Congratulations Greg and Jenny!

**February 2000**

### Hui Zhang Selected Sloan Foundation Fellow

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

Most of the this year's PDL graduate students and select faculty. Let's not ask Dave what he was thinking.

Happy 30th Birthday Greg!

# MEMS-Based Storage

*… continued from pg. 1*

cation processes, we abandon the rotating disk paradigm in favor of using simple microelectromechanical systems (MEMS) to position probe tips over the storage media.

The figure on page 1 depicts CMU's prototype MEMS-based data storage system. Similar to disk drives, the device has recording heads and a recording media surface that moves. However, the recording heads are actually MEMS probe tips that are fabricated in a parallel wafer-level manufacturing process. The CMU prototype employs magnetic storage media much like that used by disk drives. But, the media surface does not rotate; instead it translates in the X and Y directions to seek to the appropriate data. Data access is accomplished by moving the media at a constant velocity in the Y direction while data is read or written by the stationary probe tips. This design avoids problems with stiction that occur in rotating bearings at very small geometries. This is critical as stiction problems can prevent precise nanometer position control because elements tend to move by alternatively sticking and slipping. This design also avoids the potential wear (to date, MEMS bearings have tended to have quite short lifetimes) that arises when micromechanical surfaces come into contact. The media for the CMU prototype is deposited on a large (8mm x 8mm x 500um) square plate (the "media sled") and is held above the probe tip array by a network of springs. A force is applied to the sled using electrostatic actuators, though in principle electomagnetic or thermal actuators could be used. Unfortunately, such reciprocating motion is usually limited to a small fraction of the size of the structure. With typical motions being 10% or less of the suspension/actuator length, a single probe tip only "sweeps" 1% of the media sled. However, by using a large array of probe tips, all of the media area can be addressed as long as the media sled moves in X and Y by the pitch of the probe tip array. A large array of probe tips also provides a significant increase in data rate and reliability for the overall system.

Because the media surface is not perfectly flat and individual probe tip heights can vary across the probe-tip array due to both manufacturing variations and curvature of the media sled, nearly all MEMS-based storage approaches incorporate some form of tip height control. CMU's prototype provides for independent active control of the Z motion at every probe tip. Individual probe tips are placed on cantilevers that are electrostatically actuated to a fixed distance from the media surface using a local Z-positioning feedback loop.

Wiring the MEMS-based storage system's 6,400 probe tips' servo and channel electronics requires the electronics to be integrated directly into the same die as the probe tips. This integration greatly improves the bandwidth and sensitivity of the capacitive sensors that are integrated into the probe tips to determine their Z positions relative to the media. To achieve a highly-integrated CMOS+MEMS process, we have developed a series of post-processing steps following a standard CMOS fabrication that turns conventional interconnect into released movable mechanical structure. Fur-

*… continued on pg. 15*

## NEW STORAGE CONFERENCE

Professor Darrell Long, of UC Santa Cruz, has spearheaded the launch of a new storage conference called "File and Storage Technology Conference." FAST will bring together the top storage systems researchers and practitioners, providing a premiere forum for discussing the design, implementation, and use of storage systems. With a broad view of storage and file systems, FAST encourages submission across the storage spectrum including topics in data layout, caching, replication, file systems, scalable storage systems, storage security, large-scale storage applications, reliability, availability and integrity, storage management, network-attached storage and storage area networks, new storage technologies, compiler support for storage and I/O, performance evaluation, storage workloads, and QOS for storage.

FAST's steering committee includes people from industry and academia, including Darrell Long (UCSC), Dave Anderson (Seagate), Garth Gibson (CMU), John Howard (Sun), Margo Seltzer (Harvard), Kirk McKusick, Jai Menon (IBM), Merritt Jones (Mitre), and Dave Patterson (UC-Berkeley). The first FAST conference will be held in early 2002. The conference will consist of two days of technical presentations, including refereed papers, invited talks, one or two refereed industrial track session, and an introductory keynote address. Refereed papers will be published in the proceedings, provided free to technical session attendees, and available for purchase from USENIX. We are also trying to have the proceedings distributed to ACM SIGOPS members. FAST is actively seeking industrial sponsors for the conference. For more information, contact Darrell Long (darrell@cse.ucsc.edu).

**DISSERTATION ABSTRACT:**

## Scalable and Manageable Storage Systems

Carnegie Mellon University
Dissertation, September 26, 2000.

Khalil Amiri, E.C.E.

Emerging applications such as data warehousing, multimedia content distribution, electronic commerce and medical and satellite databases have substantial storage requirements that are growing at 3X to 5X per year. Such applications require scalable, highly-available and cost-effective storage systems. Traditional storage systems rely on a central controller (file server, disk array controller) to access storage and copy data between storage devices and clients which limits their scalability.

This dissertation describes an architecture, network-attached secure disks (NASD), that eliminates the single controller bottleneck allowing throughput and bandwidth of an array to scale with increasing capacity up to the largest sizes desired in practice. NASD enables direct access from client to storage devices, allowing aggregate bandwidth to scale with the number of nodes.

In a NASD system, each client acts as its own storage (RAID) controller, performing all the functions required to manage redundancy and access its data. As a result, multiple controllers can be accessing and managing shared storage devices concurrently. Without proper provisions, this concurrency can corrupt redundancy codes and cause hosts to read incorrect data. This dissertation proposes a transactional approach to ensure correctness in highly concurrent NASD arrays. It proposes distributed device-supported protocols that exploit trends towards increased device intelligence to ensure correctness while scaling well with system size.

Emerging NASD storage arrays consist of storage devices with excess cycles in their on-disk controllers, which can be used to execute filesystem function traditionally executed on the host. Programmable storage devices increase the flexibility in partitioning filesystem function between clients and storage devices. The heterogeneity in resource availability among servers, clients and network links causes optimal function partitioning to change across sites and with time. This dissertation proposes an automatic approach which allows function partitioning to be changed and optimized at runtime by relying only on the black-box monitoring of functional components and of resource availability in the storage system.

**DISSERTATION ABSTRACT:**

## A Stateless Core Approach for Scalable Internet Services

Carnegie Mellon University
Dissertation, September 25, 2000.

Ion Stoica, E.C.E.

As the Internet evolves into a global communication infrastructure, there is a growing need to support powerful and flexible services such as traffic management and quality of service (QoS). Over the past decade, two classes of solutions have emerged: those maintaining the stateless property of the original Internet architecture (e.g., Differentiated Services), and those requiring a new stateful architecture in which routers maintain per flow state (e.g., Tenet, Integrated Services). While stateless solutions are more scalable and robust, stateful solutions can provide services with higher flexibility, utilization, and assurance levels. In this dissertation, we present a novel technique and a network architecture that bridge this long-standing gap between stateless and stateful solutions. The key idea behind the technique, called Dynamic Packet State (DPS), is that, instead of having routers maintain per-flow state, packets carry this state. Based on DPS, we have developed a network architecture called Stateless Core (SCORE) in which core routers do not maintain any per-flow state. Yet, by using DPS to coordinate actions of edge and core routers along the path traversed by a flow, distributed algorithms can be designed to emulate the behavior of a broad class of stateful networks in SCORE networks. In this dissertation we describe complete solutions including architectures, algorithms and implementations which address three of the most important problems in today's Internet: providing QoS guarantees, differentiated QoS, and flow protection.

## IN APPRECIATION

The Parallel Data Lab would like to thank Intel (www.intel.com) for its recent generous donation of $40,000 for our networking research and 20 multiprocessor motherboards and 35 Pentium III CPUs that we assembled into complete workstations. Please see the web page detailing the build process at www.pdl.cs.cmu.edu/News/cases.html.

We would also like to thank Hewlett-Packard Labs for their donation of $60,000 for SCSI over IP research.

## STAFF

Patty Mackiewicz, after 5 years with the PDL and eleven with Carnegie Mellon, has moved become a Partner Liaison with the Carnegie Technology Education (CTE), a distance learning company affiliated with CMU. We all miss her very much. Fortunately, her offices are just across the street, so we can still visit.

Joan Digney, PDL's technical writer and webmaster, has moved across the continent to Medicine Hat, Alberta, Canada and is now successfully operating PDL North.

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

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

Craig Soules became a staff programmer with the PDL in January and moved on to begin work on his Ph.D. in CS this fall.

Jennifer Landefeld felt the call of Panasas in January this year and is finding entertainment there as the Operations Manager. She is now running their operations in the Redwood City, CA office.

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

## GRAD STUDENTS

Khalil Amiri defended his Ph.D dissertation on September 26 and is graduating this fall from ECE. He is moving to New York City following his defense to work with IBM. Please see page 13 for the abstract of his dissertation.

Fay Chang will also be defending her Ph.D. work and graduating from CS this fall. At this point, she is considering a move to Compaq in Palo Alto, CA.

Following the completion of his M.S., Jeff Butler left the PDL at the end of December to work as a programmer at Panasas.

Ed Hogan also left at the end of last year to program at Panasas.

Charles Hardin left the PDL at the end of December, after graduating with an M.S. in ECE to join 2Wire, a DSL and home networking development company.

Chris Sabol joined Charles at 2Wire in San Jose, CA after graduation this spring.

Mike Scheinholtz received his M.S. in ECE and now works at Mirapoint, an internet messaging network company in Sunnyvale, CA.

Ion Stoica defended his Ph.D. dissertation on September 25 and is taking a position as assistant professor at UC Berkeley in Berkeley, CA. Please see page 13 for the abstract of his dissertation.

## UNDERGRADUATES

Jeremy Praissman, a CS undergraduate, joined the PDL as a programmer in June.

Paul Cassella, a senior in CS, joined us as an undergrad programmer at the beginning of January, finishing his term of employment with the PDL in June.

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

Group photo from PDS Workshop & Retreat, November 1999

*… continued from pg. 12*

ther, extensions to this integrated CMOS+MEMS process are being developed to fabricate the read/write probe heads. Further, best use of the media requires that the media sled move by at least the probe tip actuator pitch in X and Y. Our current targets are a probe tip array with 100um centers in X and Y; hence, the media actuator must move at least 50um.

Of course, the ultimate success of MEMS-based data storage depends on its price and the performance gains in terms of speed, power, or robustness that it offers. Our simulation results show MEMS-based storage devices decrease average I/O service time an order-of-magnitude over disk drives (0.52 ms vs. 10.1 ms). This translates into large reductions in application I/O stall time (e.g., 0.3 sec. vs. 22.3 sec. on TPC-D #6). Moreover, MEMS-based storage's ability to rapidly power-down and its lower data-access power consumption creates an order-of-magnitude decrease in power consumption over a modern low-power disk drive (e.g., 350 joules vs. 6000 joules for Netscape).

Given these performance improvements, there are many opportunities for MEMS devices in the storage hierarchy. Besides replacing disks, MEMS-based storage devices could serve as a non-volatile disk cache, absorbing write traffic at a much greater speed than conventional disk drives. Further, the cache could be explicitly exposed to and managed by software, allowing software to make customized allocation decisions based on the performance needs and access patterns of various data objects, such as metadata, small files, and files with real-time constraints (e.g., video).

For many "portable" applications such as notebook PCs, PDAs, and video camcorders, MEMS-based storage provides a more robust and lower-power solution. Unlike rotating storage, which cannot cope with device rotation (e.g., rapidly turning a PDA) and is very sensitive to shock (e.g., dropping a device), MEMS-based storage is much more immune to gyroscopic effects and can absorb much greater external forces. Further, MEMS-based storage creates a new low-cost entry point for modest-capacity applications in the 1-10 GB range. This is because disks' assemblies of mechanical components keep manufacturing costs from falling below a certain point, while MEMS-based storage rides the linear decline in IC manufacturing costs. With new applications aggressively creating massive amounts of data, we are also exploring how MEMS-based data storage devices can help solve data archival problems, including capacity, time to access data, and long-term data retrieval. For example, medical imaging generates gigabytes of data per patient, which, for cost reasons, is usually stored directly on tape. Write-once MEMS devices provides an attractive alternative to tape. With areal densities 10X greater than high-capacity tape, it should be cost-effective to build storage "bricks" that hold 1000s of MEMS devices. Each brick would hold Petabytes of data that could be accessed in under 1 second. Further, by incorporating logic into the MEMS-based storage device, it would be possible to process data directly within the storage brick. With massive numbers of storage bricks there is massive computational parallelism available, creating the ultimate active disk.

Another application domain for MEMS-based storage is bulk non-volatile storage for embedded computers. Single-chip "throw-away" devices that store very large datasets can be built for such applications as civil infrastructure monitoring (e.g., bridges, walls, roadways), weather or seismic tracking, and medical applications. For example, one forthcoming application is temporary storage for microsatellites in very low earth orbit. Given that a satellite in a very low orbit moves very quickly, communications are only possible in very short bursts. Therefore, a low-volume, high-capacity, non-volatile storage device is required to buffer data. MEMS-based storage devices could also add huge databases to single-chip continuous speech recognition systems and be integrated into low-cost consumer or mobile devices. Such chips could be completely self-contained, with hundreds of megabytes of speech data, custom recognition hardware, and only minimal connections for power and I/O.

CHIPS envisions a bright future for MEMS-based storage technologies and their role in enabling a complete system-on-a-chip solution. More information on CHIPS is available at www.chips.ece.cmu.edu. A detailed description of our MEMS-based Storage Device is in the November, 2000 issue of the Communications of the ACM.

The real brains behind the whole operation.

*… continued from pg. 9*

to hybrid approaches that leverage the best of semiconductor memories and disk drives. From semiconductor memories, the hybrid approaches adopt the parallel wafer fabrication process which keeps unit costs low. From disk drives, the hybrid approaches adopt recording heads that use mechanical position to address data stored in a thin film material instead of the lithographic definition of address required by today's semiconductor memories. But, for compatibility with silicon wafer fabrication processes, these hybrid approaches abandon the rotating disk paradigm in favor of using simple microelectromechanical systems (MEMS) to position probe tips over the storage media.

MEMS-based storage systems have the potential to create a whole new storage technology capable of achieving a quantum decrease in entry cost, access time, volume, mass, power dissipation, failure rate, and shock sensitivity. Perhaps more importantly, these devices can integrate computation with storage, creating complete system-on-a-chip solutions – including mass storage. This will enable many new applications exploiting the low unit cost and extremely small size of these new hybrid devices; e.g., "intelligent" appliances, sophisticated teaching toys, biomedical monitoring devices, civil infrastructure monitoring devices, micro- and nano-satellites, highly-integrated archival storage systems, highly-secure systems, etc. This is not just future fiction – the technologies needed to build these hybrid devices are already emerging, making it likely that a broad market for non-volatile rewritable mass storage devices will develop within the next five years.

This paper examines MEMS-based data storage technology and how it can be applied to computer systems using a prototype MEMS-based data storage system currently under development at Carnegie Mellon University (CMU) as a design example. When possible, we will compare and contrast the CMU MEMS-based storage system with others currently under development at industrial research laboratories such as IBM, HP and Kionix.

## Survivable Information Storage Systems

*Wylie, Bigrigg, Strunk, Ganger, Kiliccote & Khosla*

IEEE Computer, August 2000.

ABSTRACT

As society increasingly relies on digitally stored and accessed information, supporting the availability, integrity and confidentiality of this information is crucial. We need systems in which users can securely store critical information, ensuring that it persists, is continuously accessible, cannot be destroyed and is kept confidential. A survivable storage system would provide these guarantees over time and despite malicious compromises of storage node subsets. The PASIS architecture combines decentralized storage system technologies, data redundancy and encoding, and dynamic self-maintenance to create survivable information storage.

## Network Attached Storage

*Gibson & Van Meter*

Communications of the ACM, Vol. 43, No. 11 November, 2000. Special issue on: Ultra-High-Density Data Storage.

ABSTRACT

The rapidly growing market for networked storage is responding to the exploding demand for storage capacity in today's increasingly short-staffed and Internet-dependent world. Storage area networks (SAN) and network attached storage (NAS) are two proven approaches to networking storage. Technically, the presence of a file system in a storage subsystem differentiates NAS, in which it is present, from SAN, in which it is absent. In practise, however, it is often the binding of NAS to Ethernet network hardware and SAN to Fibre Channel network hardware that has the greatest impact on a customer's system. This article is about the changes in technology that may blur this network-centric distinction between NAS and SAN. For example, the decreasing specialization of SAN protocols promises SAN-like devices on Ethernet network hardware. Alternatively the increasing specialization of NAS systems may embed much of the file system into storage devices themselves.