PARALLEL DATA LABORATORY
CARNEGIE MELLON UNIVERSITY

THE
PDL Packet
THE NEWSLETTER ON PARALLEL DATA SYSTEMS • FALL 1995

I•N•S•I•D•E

YEAR•IN•REVIEW

September 94	
PDIS94	Hugo Patterson, PDL, presented TIP (Mach 3.0)
October 94	
CMU	Roger Wood, IBM, talked on trends in storage
CMU	Rick Rashid, Microsoft, talked on Tiger video server
CMU	DSSC Annual Review
November 94	
CMU	PDL retreat and workshop
CMU	ARPA site visit for Multicomputing project
CMU	Garth Gibson, PDL, talked about parallel file systems

... continued on p.10

ARPA Set to Fund NASD Research Initiative

The Advanced Research Projects Agency (ARPA) plans to approve Carnegie Mellon University's proposal to research, develop, and build network-attached, secure disks (NASD). The proposal, submitted in response to a broad call from ARPA for reinventing OS service structure, outlines work that is aimed at eliminating the workstation server as a bottleneck from networked file systems by promoting the disk to a first-class network device. Once formally approved, CMU researchers—primarily members of the Parallel Data Laboratory (PDL)—will start research into secure file system, drive software, and hardware design.

Under the NASD proposal, portions of the network file systems and security will be embedded into the disk drive. By embedding these systems, researchers aim to minimize latency between disk drives and the client. With integrated security, NASD will be better able to provide networks with an integrity that is independent of the trustworthiness of client operating systems.

If they successfully raise network security and integrity while lowering the amount of time it takes to move data, NASD disks will be able to serve as building blocks for file systems that are highly reliable, highly available, and fault tolerant.

Critical to the success of the NASD project is an industry-standard interface for the network-attached disk drives. Currently, researchers in the PDL are collaborating with a working group in the National Storage Industry Corporation (NSIC) to define the network attached storage interface. Currently the plan is for the NASD storage interface to include a two-level scheme (object, offset) for storage addressing and a lightweight private-key authentication and encryption protocol. NASD and NSIC collaborators are expected to meet every few months. Announcements related to this working group are distributed on the mailing list *nasd@sun.com*.

> **Needless to say, I'm pleased.**
>
> **—Garth Gibson, principal investigator for the NASD project**

In broadly researching secure, scalable storage architectures, CMU researchers will build on research currently supported by

... continued on p.4

Hot TIP: DASD Evolves into NASD

From the Director's Chair

GARTH GIBSON

I'm pleased to report that the second full year of the Parallel Data Laboratory has been a good one. We've achieved crucial successes, from my perspective, in our filesystems implementation and in developing our new research directions. Along the way, we have advanced our other projects systematically, done the academic thing with regularity, and expanded our experimental infrastructure as quickly as our nerves can manage.

This year, under Hugo Patterson's leadership, the Informed Prefetching and Caching project has reworked the entire buffer cache of Digital's OSF/1 version 2.0 was reworked to aggressively act on application-provided hints. This informed resource management, which was applied to six I/O-intensive applications—full-text search, 3-D visualization, speech recognition, object code linking, computational physics, and database joins—demonstrated reductions in elapsed time of 20% to 85%. Drawing mainly on informed prefetching's ability to increase the number of concurrent I/Os issued into a parallel storage system. However, informed caching can also deliver large reductions in elapsed time by increasing the buffer cache's hit ratio for applications like database and object linking, particularly when storage concurrency is limited. I'm especially pleased to report that our paper describing this work will appear in the premier OS conference, the Symposium on Operating Systems Principles (SOSP), Dec 3-6 this year. More on TIP later in this newsletter.

Our new research directions are the other major event this year. The new effort, Network-Attached Secure Disks (NASD). NASD is all about the evolution of the disk-drive-embedded controller and its communications infrastructure. It will draw on our RAID experience with failure tolerance and high availability. It will draw on our parallel-file-system research by providing low-level file system support in the disk drive appropriate for distributed and parallel file systems. And most satisfyingly, it will close the circle on our informed prefetching work. This may seem counterintuitive, but in fact Hugo and I started out to work on new smart-controller designs three years ago. At that time we decided that before controller smarts would be particularly effective, the system that used them would need to provide much more advance notice of the work it needed done; hence, we developed strategies for aggressive prefetching. With NASD, we return to controller design and add a new focus area in NASD: security. With a network-attached disk, unauthorized machines and users can send commands with impunity. NASD will develop security protocols to fend off this threat, yielding distributed file systems whose integrity is secure against malicious network attacks and resilient when its clients operating systems are compromised.

While these two accomplishments were more than enough for me, PDL members more interested in RAID have not been idle. RAIDframe, a rapid prototyping system for RAID architectures, was envisioned, designed, constructed, refashioned into kernel, user-level, and simulation instantiations, and will soon be released for your use. Although intended largely as a new, more functional evaluation platform for our RAID architecture research, RAIDframe enables us to treat RAID archi-

... *continued on p.6*

THE PDL PACKET

The Parallel Data Laboratory
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15217-3891
412•268•3835 VOICE
412•268•3010 FAX
412•268•5576 FAX ALTERNATE

PUBLISHER
Garth Gibson

EDITOR AND DESIGNER
LeAnn Neal

The *PDL Packet* is published one or more times a year and mailed to members of the Parallel Data Consortium. Copies are given to other researchers in industry and academia as well. Postscript versions will reside in the Library section of the PDL Web pages. 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.

PARALLEL DATA LABORATORY

MISSION OF THE LAB

To advance the state of the art in storage subsystems and to integrate them efficiently into parallel file systems, high bandwidth networks, and multicomputers.

CONTACTING US

WEB PAGES
http://www.cs.cmu.edu/Web/Groups/PDL

DIRECTOR
Garth Gibson
412•268•5890
garth.gibson@cs.cmu.edu

STAFF MEMBERS
Chris Demetriou
chris.demetriou@cs.cmu.edu

Patty Mackiewicz
patty.mackiewicz@cs.cmu.edu

LeAnn Neal
leann.neal@cs.cmu.edu

Jim Zelenka
jim.zelenka@cs.cmu.edu

GRADUATE STUDENTS
Khalil Amiri
khalil.amiri@cs.cmu.edu

Fay Chang
fay.chang@cs.cmu.edu

Bill Courtright
bill.courtright@cs.cmu.edu

Eka Ginting
eka.ginting@cs.cmu.edu

Qingming Ma
qingming.ma@cs.cmu.edu

Hugo Patterson
r.hugo.patterson@cs.cmu.edu

Erik Riedel
erik.riedel@cs.cmu.edu

Daniel Stodolsky
daniel.stodolsky@cs.cmu.edu

Major New Development: RAIDframe, A Tool for Developing RAID Systems

by LeAnn Neal

A team of researchers in the Paralleld Data Lab is currently developing an extensible framework for experimenting with RAID architectures. RAIDframe, which uses directed acyclic graphs (DAGs) to model RAID operations, allows array designers to implement and evaluate new RAID architectures quickly and accurately.

Because it has a DAG-based structure, RAIDframe also addresses a particularly thorny issue for most RAID manufacturers: error handling. In RAIDframe, recovery is being automated by embedding the criteria for recovery action into DAG execution.

RAIDframe consists of an execution engine, a library of primitive RAID operations and DAGs, DAG selection code, disk queues, cache, and an address-mapping module.

A number of RAID architectures are being implemented using RAIDframe. They are: RAID Levels 0, 4, 5, and 6 as well as parity declustering for RAID Level 5, declustering with distributed sparing, parity logging, write deferring, and log-structured storage.

RAIDframe runs at three levels: a user-level software array controller; a simulator; and a device driver in the kernel. Depending upon their needs, RAID designers can test new designs against simulated disks, or real disks directly or with a trace-driven process. The software is currently supported for DEC's OSF/1 versions 2.0 and 3.2.

Because the goal is wide distribution to RAID developers, documentation is being written which describes the motivation, background, and technical details for RAIDframe. Rules for defining primitives and creating DAGs are specified to guarantee extensibility and flexibility of the development environment. Details for installing, using, and extending the system wrap up the document, which will be available soon.

Workshop and Retreat Scheduled for Oct. 30-Nov. 1

The annual PDL workshop and retreat is scheduled for October 30 through November 1. As in the past, industry sponsors from the Parallel Data Consortium as well as academic researchers from the CMU community will be invited for three days of intensive review and collaboration.

Unlike previous workshops, this year's workshop is being held independently of the fall DSSC review, which is September 19-20. Scheduling the workshop later in the fall will allow PDL researchers to present results more effectively from two projects: RAIDframe and informed prefetching and caching.

Last year's workshop featured talks on RAID (error recovery, declustering, deferring writes, mobile disk arrays), file systems (cache management, asynchronous name resolution, and compiler-generated hints for prefetching), and parallel programming research.

Other topics for this year's talks include: Parallel Flows Networking Service, Scotch Parallel File System, Network-Attached Secure Disks, and the Scalable I/O Application Programming Interface.

This year's retreat will be held at the Wisp Resort in Deep Creek Lake, MD. A 3-hour social hike is being planned along with the talks. Consortium members are encouraged to send one or two people to the workshop.

... continued from p 1

ARPA to Fund NASD Initiative

ARPA under its Software Systems for High-Performance Multicomputing. For example, researchers plan to demonstrate the proposed security protocol by using the Scotch Parallel File System (SPFS) along with something close to the Kerberos authentication service. In addition, they are working with ARPA's Scalable I/O Initiative to ensure that SPFS complies with the Initiative's low-level application programming interface (API) for I/O-intensive parallel programs.

Another example of current ARPA-funded research that will plug into the NASD effort is the informed prefetching and caching project. In this case, researchers will use results from this project to demonstrate how the two-level storage system interface enables device-specific intelligence.

Besides approving the base research contract, ARPA is expected to support the proposed option to design a distributed video service using NASD storage devices. It is hoped that this will strengthen the Informedia project, a joint effort between CMU and the award-winning public television station WQED. **Please see sidebar below.**

ARPA is expected to fund the NASD proposal later this year. For anyone interested, copies of an early white paper can be retreived via anonymous ftp from ftp.cs.cmu.edu in directory /afs/cs/project/pdl/ftp/NASD.

RAID Tutorial Available at PDL Web Site

Prof. Gibson gave a tutorial on Storage Architectures: RAID and Beyond at both SIGMOD95 and ISCA95.The tutorial covered these topics

- RAID basics: striping RAID levels, controllers
- Recent advances in disk technology
- Expanding RAID markets
- RAID reliability: high and higher
- RAID performance: fast recovery, small writes
- Exploiting disk parallelism: deep prefetching
- RAID over the network

The tutorial was converted to HTML for viewing on the Web and is available from the PDL Web site listed in the masthead on page 3.

Informedia Digital Video Library: Interactive, Networked Video on Demand

The National Science Foundation, Advanced Research Projects Agency, and the National Aeronautical Space Administration together have granted Carnegie Mellon University $4.8 million over four years to develop an on-line, interactive, digital video library that will enhance the study of science and mathematics for students in kindergarten through college.The Informedia project, to be created by CMU and public television station WQED/Pittsburgh, will integrate speech, image, and natural language understanding technologies developed by university researchers to access, explore, and retreive video material from the archives of public television and educational institutions.

Initially, the library will contain about 1,000 hours of raw and edited video from the archives of WQED, video developed by the Fairfax County, VA, public schools, and video course material produced the British Broadcasting Corp. for the Open University, Leeds, England, a college without walls with an enrollment of more than 200,000 students.

Digital Equipment Corp. will supply the project's hardware, including site servers and user desktops.The company is contributing hardware that spans a range of its products, from PCs through sophisticated Alpha AXP-based workstations and video servers. Microsoft Corp. will contribute software and financial assistance to the project. Bell Atlantic Corp. will provide funding for the system's communication services.

As computer scientists create the library and tools for exploring it, they will incorporate emerging standards in storage, high-bandwidth communications and services as well as in displaying and manipulating data.To handle storage and delivery, Informedia will initially require one terabyte of storage and a metropolitan area network with switched, multimegabit servers. Because researchers are concerned with the need to sustain sufficient data rates from the file system and over the network in order to provide high quality video and audio to library users, they have begun collaborating with the NASD project (see main story above) to study the needed distributed-video-storage capabilities.

To handle network billing, access control, security and privacy, Informedia will draw upon research from CMU's NetBill project.These features will help ensure that copyright owners are properly compensated for their work, thereby enticing them to contribute to the video collection. They will also ensure that all users are charged appropriately for the video they access. Access control will allow the system to specify which users can access which services.

Faculty Collaborate with PDL on NASD

In addition to our own beloved Garth Gibson who will act as the project's technical contact for ARPA, four other faculty researchers are slated to work on the network-attached, autonomous, secure disks. On the following two pages, each of these researchers is profiled.

Ronald Bianchini

Associate Professor
Electrical and Computer Engineering and Computer Science

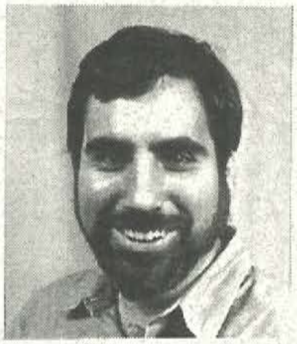
Professor Bianchini's research interests include computer architecture, computer networks, distributed systems, and telecommunications switching.

Professor Bianchini is involved in the design of a novel architecture, Tera, for switching packets in telecommunictions networks. Tera consists of a shared-memory input queue that permits simultaneous access by all switch input ports and non-blocking output ports. For N inputs and N outputs, Tera consists of a shared memory with N interleaved banks and multiple Banyan networks with complexity NlogN. The major advantages of the Tera architecture are its robust performance under bursty and unbalanced traffics and its scalability to large networks. Tera is shown to have multiple orders of magnitude fewer dropped packets than other proposed packet switches and to reduce communiction latency significantly. Prototypes of the switch for ATM traffic are planned with up to 32 input and output ports. The current focus of this work is the deployment of the Tera switch in the ECE LAN environment, including encapsulation of Ethernet traffica via ATM.

Another research focus has resulted in the first practical application and implementation of distributed systems-level diagnosis theory which specifies distributed-system fault tol erance. In that specification, each of several units in a distributed system attempt to determine the fault conditions, or diagnosis, of the remaining units in the system. Although there has been significant reseach into this theory that remains unimplemented, it has resulted in numerous practical algorithms. The first, Adaptive DSD, is proven to require the minimum possible overhead by adapting resource use to the current fault situation. The algorithm is currently implemented on 200 Unix workstations of the ECE department and has executed continuously for over two years even though no single workstation has remained fault free for the entire period.

Doug Tygar

Associate Professor
Computer Science



Professor Tygar's research is in supporting electronic commerce using tools from distributed systems, computer security, cryptography, protocol theory, applied algorithms, and applied economics.

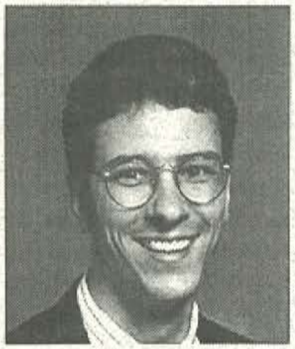
He is engaged in three other projects: Dyad, NetBill, and Electronic Franking.

The Dyad project is building secure coprocessors that use physical mechanisms to protect memory. If the processor is breached, then all memory is erased. Dyad's design has prompted five different vendors to announce their own secure coprocessors. Project members have ported an operating system onto the secure coprocessor which can run secure remote execution routines. Based on this work, Visa International recently announced plans to convert existing credit cards into smart cards.

The NetBill project is building a billing server that can be used to charge information purchases over the Internet. Unlike credit cards which have a marginal transaction cost of twenty-five to sixty cents per purchase, NetBill has a marginal transaction cost of under one cent. NSF and ARPA have selected NetBill to support electronic commerce for CMU's Informedia project. Fundamental research questions include: How can various pricing schemes be supported? How can sealed proof of a transaction be provided? How can certified delivery be guaranteed so customers are charged exactly when they receive material? Is it possible to mark documents with low-level dithering codes that uniquely identify the documents and discourage illegal pirating of copyrighted material?

The Electronic Franking project has been investigating new standards involving cryptographic methods and 2-D bar codes for computer-generated postage indicia. The U.S. Postal Service has announced its intention to use a variant of the project's standards, starting in a year. Based on this work, Canada Post and (UK) Royal Mail are also moving towards similar standards. It is the goal of this work to support more general forms of electronic commerce than merely supporting mail.

CS and ECE Researchers Bring Expertise



Dave Nagle

Assistant Professor
Electrical and Computer Engineering

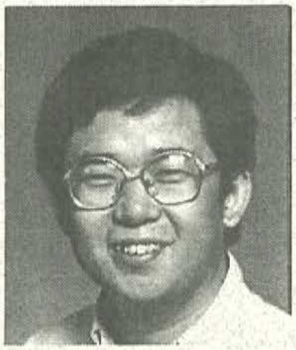
Professor Nagle's research interests include computer architecture, operating systems, and software/hardware interface design.

Prof. Nagle spent the past few years examining the influence operating system software has on architectural design. Using a range of UNIX-based operating systems—each with radically different internal organizations—and a basic RISC-workstation architecture as an experimental base, he extended quantitative-analysis techniques to the study of OS/architecture interactions.

As a result of this research, he found that operating systems such as OSF/1 and Mach 3.0 suffer significant performance penalties,increasing application runtime up to 100% over a more mature operating system, Ultrix. Further, he found that nearly 40% of this loss was due to unforeseen interactions between the operating system and the architeture. Based on these findings, he identified a number of issues that architects should consider in order to design better support for operating system and software technologies.

Prof. Nagle is also a co-developer of the trap-driven simulation technique which allows workstations to perform very fast, memory-system simulation.The trap-driven simulation technique is currently being extended by Digital Equipment Corporation for use in the design of their next generation Alpha processor. Prof. Nagle has spent the summer consulting at DEC with the advanced microprocessor group.

Complementing this research is his work in low-power architecture design.The goal of this work is to use Carnegie Mellon's low-power circuit technology to build ultra-low power processors.A central problem in this work is managing the flow of data to minimize the power consumed while providing good performance.



Hui Zhang

Assistant Professor
Computer Science and Electrical and Computer Engineering

Prof. Zhang's research interests span the theory, design, and implementation of scalable and efficient integrated-services computer networks.

For the last few years, he has worked on resource management algorithms and protocols to support real-time communication in packet-switched networks, specifically on U.C. Berkeley's Tenet Real-Time Protocol Suite, which guarantees real-time service across heterogeneous wide-area networks. Currently, he is researching support for delivering bursty, compressed video in real time while maintaining high network use.

Prof. Zhang is also developing algorithms that support multi-point communications for applications that can tolerate limited packet losses, such as video conferencing, and those that need reliable transport, such as distributed simulations, shared whiteboard, news and information dissemination. For the first type, he worked with researchers at Berkeley to design and implement a video gateway architecture that allows heterogeneous receivers with access to networks of different speeds to participate in the same video conference. For multi-point applications that need reliable transport, he is developing an efficient, scalable, reliable, multicast protocol that supports timely dissemination of information to a large number of receivers in a heterogeneous network. Key problems to be solved for multi-point networking are the signaling protocol, the routing algorithm, and the resource management algorithm.

... *continued from p.2*

Hot TIP: DASD Evolves into NASD

tectures as "programs" in its directed acyclic language, providing automated error recovery and the opportunity to apply automatic optimization.

Moreover, we have maintained our academic posture by reporting our work to four conferences (PDIS, CMG, COMPCON, and SOSP), educating computer architects (at ISCA) and database researchers (at SIGMOD) with a storage technology tutorial, and defending three Ph.D. proposals (Bill, Qingming, and Danner) and a master's thesis (Rachad).

Finally, in our spare time (huh?), we have enlarged our research testbed infrastructure with an OC3 (155 Mbits/sec.) ATM networking.

So there you have it. Or rather, I've given you a sampling of what we've been up to this year. Most of these topics are covered in greater detail elsewhere in this newsletter, but the best way to get more information is to get yourself invited to our impending workshop and retreat, Oct 30 - Nov 1.

Read on to find out how to do that.

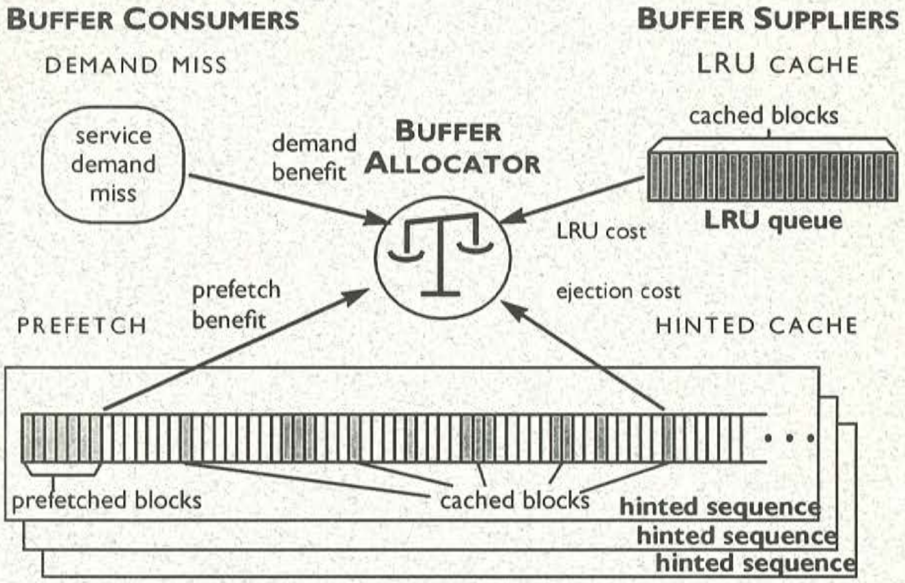
Informed Prefetching and Caching to be Presented at SOSP

The PDL paper "Informed Prefetching and Caching" will be presented at the Fifteenth ACM Symposium on Operating Systems Principles (SOSP) at the Copper Mountain Resort, CO, Dec. 3-6, 1995. The paper presents aggressive, proactive mechanisms that tailor file system resource management to the needs of I/O-intensive applications. These mechanisms consist of application disclosures (hints) about future accesses and an informed cache manager which dynamically allocates file cache buffers.

Hints can be used to prefetch data, thereby masking disk latency or to cache already used data to avoid disk access altogether. However, using hints creates competition between the traditional LRU queue and the hinted cache for a limited resource: file buffers. The paper presents a cost-benefit analysis for resolving the tension between supply and demand and describes a cache manager which uses this analysis.

The figure at the right shows this continual balancing act between buffer consumers and buffer suppliers. The cache manager needs to know whether a cache buffer should be used to demand fetch or prefetch data now, and, if so, which block should be ejected to free a buffer. To answer these questions, each potential buffer supplier or consumer has an estimator that independently computes the value of its use of a buffer. The buffer allocator uses these estimates (which are standardized) to reallocate buffers to reduce overall I/O service time.

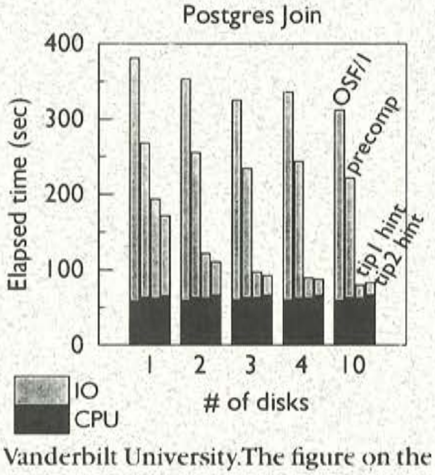
We implemented informed prefetching and caching in Digital's OSF/1 operating system and measured its performance on a 150 MHz Alpha equipped with 15 disks running a range of applications. Informed prefetching reduces the execution time of text search, scientific visualization, relational database queries, speech recognition, and object linking by 20-83%. Informed caching reduces the execution time of computational physics by up to 42% and contributes to the performance improvement of the object linker and the database. Moreover, applied to multiprogrammed, I/O-intensive workloads, informed prefetching and caching increase overall throughput.

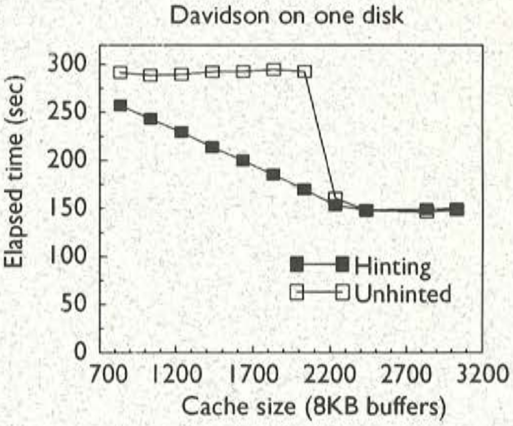


Benefit of Informed Caching for Single Applications

One application that we tested is Postgres, a relational database from the University of California at Berkeley. In our test, Postgres executes a join of two relations. The figure on the left shows the elapsed time for the join in the standard version of Postgres, a restructured version that precomputes offsets for the inner relation, and in the restructured Postgres when it gives hints. Restructuring improves access locality and cache performance, allowing it to run faster than standard Postgres. Delivering hints then greatly reduces I/O stall time.

Another application we tested is the Davidson algorithm, a computational physics program developed by Vanderbilt University. The figure on the right shows that informed caching in TIP-2 discovers an MRU-like policy which uses additional buffers to increase cache hits and reduce execution time. In contrast, LRU caching derives no benefit from additional buffers until there are enough of them to cache the entire dataset, which is 16.3 MB (2089 8K blocks).





Scalable I/O Initiative

This year the PDL has expanded its role in the Scalable I/O Initiative, a nation-wide community of academics and practitioners of parallel machines, applications, languages, tools, and operating systems funded by the NSF, ARPA, and NASA.

Originally signing up only to explore disclosure-based prefetching for parallel file systems, PDL participants expressed an opinion on the structure of portable, high-performance parallel file systems; they proposed a two-level structure. At the high level, programmers should be provided with a variety of application-specific, structure-rich interfaces to improve ease of use and expressive power. Separated into the lower level should be all the performance enabling features necessary for high-performance environments: asynchronous operations, hints, scatter-gather, pass-by-reference, third-party transfer and deviceoptimizations. In this way, a low-level parallel file system, customized to a particular machine and programming environment, can support a variety of application-selected, high-level, parallel file system personalities.

PDL members are now deep in the midst of the time-critical task of defining, in collaboration with the SIO OS working group (IBM, Intel, Princeton, U. of Arizona, U. of Washington, CMU), the interface between high and low levels of such a parallel file system.

Thesis Proposals

Since the last PDL newsletter, three graduate students have proposed thesis topics and been accepted. The abstracts for each proposal follow.

Error Recovery in Redundant Disk Arrays proposed by William Courtright

Redundant disk arrays are a popular method of providing high-performance disk storage, capable of withstanding disk failures without loss of data. Since disk faults are manifested as errors, disk-array controllers are required to perform error recovery, removing the effects of the error and continuing service without interruption. Work in the array is performed concurrently and there are a large number of scenarios in which these errors may occur. Traditional approaches to error recovery either attempt to enumerate each of these scenarios, a manual task which is prone to mistakes and specific to a particular design, or simplify recovery by trading resources and performance for complexity reduction. We believe it is possible to simplify error recovery in redundant disk arrrays without the introduction of this overhead. We propose an approach, which we call compensating methods, to this end. Furthermore, we introduce the idea of modeling operatings as antecendence graphs, facilitating extensible design, correctness reasoning, and mechanized recovery. A research plan to validate this approach is included.

Parallel Flows: A New Networking Service proposed by Qingming Ma

Applications in both the scientific computing and multimedia domain process large amounts of data that are often retrieved from remote high-performance storage systems. Parallel computations over a workstation cluster may also involve large amounts of data exchange among multiple computational units. This results in network bandwidth requirements that are higher than the physical-link bandwidth in today's high-speed networks.

With the emerging switch-based networking techology (e.g., ATM, HiPPI, switched-FDDI, and switched-Fibre-channel), the aggregate bandwidth in the network greatly exceeds the link bandwidth. Since both the systems used for computation and storage typically have multiple network connections, high network throughput can be potentially achieved by making use of multiple paths available in the switch-based network.

A promising way to meet the network bandwidth requirements of applications is to split one logical data stream and transmit it through multi–ple paths in parallel. However, existing and proposed network protocols are mainly targeted for "single stream" communications. There is a lack of a network service that supports efficient data movement of parallel flows.

I propose to design and implement a new network service—parallel flows—that supports efficient parallel data movement across networks. Two novel features lead to efficiency. First, applications participate in bandwidth planning and interact with the network through an application programming interface (API). Second, the network coordinates parallel transfers through a new routing mechanism—c oordinated routing. The benefits of using this new service will be demonstrated by applications that move data across Nectar between multiple workstation nodes and a parallel file server—the Scotch Parallel File System (SPFS).

File System Support for Parallel I/O in Multicomputers proposed by Daniel Stodolsky

Multicomputers, a collection of workstations connected by a high-bandwidth network, are becoming the vehicle of choice for many classes of parallel computations. The economies of scale of the workstation market make multicomputers more cost-effective for many classes of applications than traditional vector supercomputers or massively parallel processors (MPPs), and the advent of cost-effective, high-bandwidth switched networks has removed network bandwidth as a bottleneck for communication-intensive applications.

... continued on p.9

... *continued from p.8*

Unfortunately, the lack of a scalable file system that delivers high bandwidth to parallel applications has prevented multicomputer use for I/O-intensive problems. In addition to high bandwidth, such a scalable file system must efficiently support concurrent write-sharing and be highly available to meet the I/O requirements of long-running parallel applications.

The thesis of this work is that four ideas are crucial in supporting I/O-intensive parallel applications on a multicomputer: per-file striping, application-computed reliability, application-initiated prefetching, and application-exposed data coherence. Per-file striping, the distribution of file data on storage servers which export a two-dimensional address space, eliminates shared metadata as a bottleneck and shared parity as a security threat. Application-computed reliability, the direct computation of check data by the parallel application, allows exploitation of the full-stripe write optimization to reduce the bandwidth overhead of fault-tolerant writes to much less than 100% and provide scalable computing power for reconstruction. Application-initiated prefetching, in which a parallel application describes its access pattern and the file system transparently prefetches data into the application address space, effectively hides the high latency of storage server access for the finely interleaved data-sharing patterns in many parallel programs. And last, application-exposed coherence, the providing of explicit file data consistency mechanisms to the application, removes the inefficiency of a distributed shared memory implementation in the file system with a negligible burden on the application writer. A multicomputer file system based on these four ideas can provide high bandwidth, easy data sharing, and be tolerant to failures of disks and storage nodes. This thesis will be experimentally validated by the construction of the Scotch Parallel File System (SPFS), a scalable, fault-tolerant, network parallel file system.

Master's Thesis

The abstract for Rachad Youssef's master's thesis, *RAID for Mobile Computers*, follows.

The requirement for high-performance, highly available storage for file servers and supercomputing systems led to the development of Redundant Arrays of Inexpensive Disks (RAID) and log-structured file systems (LFS). For mobile computers, however, performance is often a secondary requirement to long battery life. This study examines the design issues of low-power, highly available disk arrays for mobile computers. Specifically, by dynamically remapping the location of newly written data using a log-based allocation strategy and by deferring parity updates in an NV-RAM cache, the rate of drive spin-ups can be reduced by a factor of 2.

Lab Status

Since the last newsletter—nearly a year ago—a lot has changed in the status of the lab and its equipment. Below is a quick rundown of the major additions and changes.

Disk Drives and Arrays

Seagate donated 160 disk drives to be used in a field reliability test; these drives were canisterized into 10 disk arrays and integrated into the Scotch-3 experimental, parallel storage testbed. The SW800 cabinet is now populated with 50 HP 2247 disk drives on 16 shelves.

Machines

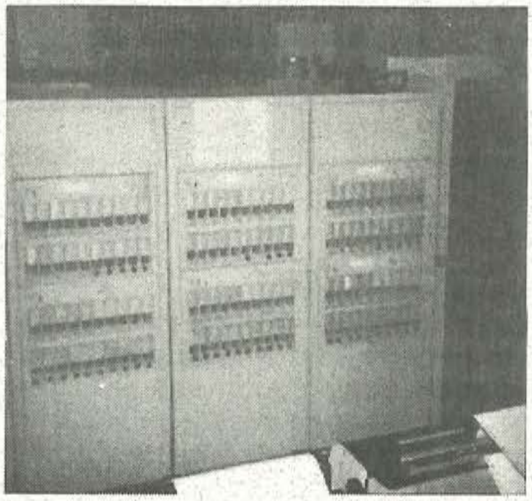
Currently, there are 10 DEC 3000 workstations in the Lab,each running DEC's OSF/1 v.3.2 and acting as parallel file servers for Scotch-3. Each server has one fast, wide, differential SCSI adapters— KZTSA models—to communicate with the ADP 93 controllers for each disk array. The Scotch servers are also equipped with ATM adapters and half of the servers have switched HIPPI adapters, too. Early this year, the Lab received nine RS6000s, IBM's new 80mhz PowerPCs which will soon be equipped with ATM. Balblair, the Lab's HP 755 workstation, is also equipped with ATM.

Networking

A FORE Systems ATM switch, which supports 16 machines, has been installed with a second switch arriving in 60 days. In addition to the Scotch

... *continued on p.11*

Parallel Data Lab Disk Arrays Are Up and Deployed in Lab, CS Machine Room



The PDL has 10 Symbios Logic 6299 arrays shown housed in 6256 cabinets up and deployed. Disks used in the arrays are two Seagate models: ST12400N and ST31200N. Four of the five cabinets are located in the machine room of the School of Computer Science (shown here), a refrigerated room attended 24 hours a day that houses other class-B equipment.

Recent Publications

The Scotch Parallel Storage Systems

In *Proceedings of the IEEE CompCon Conference*, March 5-8, 1995

Abstract

To meet the bandwidth needs of modern computer systems, parallel storage systems are evolving beyond RAID levels 1 through 5. The Parallel Data Lab at Carnegie Mellon University has constructed three Scotch parallel storage testbeds to explore and evaluate five directions in RAID evolution: first, the development of new RAID architectures to reduce the cost/performance penalty of maintaining redundant data; second, an extensible software framework for rapid prototyping of new architectures; third, mechanisms to reduce the complexity of and automate error-handling in RAID subsystems; fourth, a file system extension that allows serial programs to exploit parallel storage; and lastly, a parallel file system that extends the RAID advantages to distributed, parallel computing environments. This paper describes these five RAID evolutions and the testbeds in which they are being implemented and evaluated.

Informed Prefetching and Caching

To appear in *Proceedings of the Fifteenth Symposium on Operating System Principles (SOSP)*, Dec. 3-6, 1995

Abstract

In this paper, we present aggressive, proactive mechanisms that tailor file system resource management to the needs of I/O-intensive applications. In particular, we show how to use application-disclosed access patterns (hints) to expose and exploit I/O parallelism, and to dynamically allocate file buffers among three competing demands: prefetching hinted blocks, caching hinted blocks for reuse, and caching recently used data for unhinted accesses. Our approach estimates the impact of alternative buffer allocations on application execution time and applies cost-benefit analysis to allocate buffers where they will have the greatest impact. We have implemented informed prefetching and caching in Digital's OSF/1 operating system and measured its performance on a 150 MHz Alpha equipped with 15 disks running a range of applications. Informed prefetching reduces the execution time of text search, scientific visualization, relational database queries, speech recognition, and object linking by 20-83%. Informed caching reduces the execution time of computational physics by up to 42% and contributes to the performance improvement of the object linker and the database. Moreover, applied to multiprogrammed, I/O-intensive workloads, informed prefetching and caching increase overall throughput.

Year-in-Review Continued

December 94	
CMG95	Bill Courtright, PDL, presented error handling in RAID
SIO	Scalable I/O kick-off meeting
January 95	
CMU	Qingming Ma, PDL, proposed a thesis topic (see news briefs)
February 95	
CMU	NSF site visit to DSSC
CMU	Bill Courtright, PDL, proposed a thesis topic (see news briefs)
March 95	
CMU	Peter Corbett, IBM, talked on the VESTA parallel file system
CMU	Alok Choudhary, Syracuse, talked on parallel file system libraries
Compcon95	Garth Gibson, PDL, reported PDL research
NSIC	Working group proposed NASD (Network-Attached, Secured Disk Drives
April 95	
RAID'95	Garth Gibson, PDL, talked on RAID in the next century
May 95	
CMU	Tom Cormen, Dartmouth, talked on language support for parallel I/O
SIGMOD95	Garth Gibson, PDL, gave a tutorial on storage architectures
NSIC	First working group met on NASD
CMU	Daniel Stodolsky, PDL, proposed a thesis topic
June 95	
ISCA95	Garth Gibson, PDL, gave a tutorial on storage architectures
July 95	
Gordon Conference	Garth Gibson, PDL, presented NASD
August 95	
CMU	TIP team submitted camera-ready copy for SOSP15
CMU	Rachad Youssef, PDL, defends masters' thesis (see news briefs)
September 95	
NSIC	Second working group met on NASD
October 95	
ARPA	NASD ARPA contract scheduled to begin

... continued from p.9

servers, two RS6000s, and Balblair, a switch-to-switch connection with the ATM switch used by the Nectar gigabit testbed is running. Once the second switch arrives, seven more machines will be equipped with ATM and the two PDL ATM switches will be interconnected.

Projects and Testbeds

The Scotch storage testbeds have evolved greatly over the past year. Scotch-2, a software-managed disk array used by TIP, is installed. Scotch-3, storage for the Multicomputer project, is currently at version three with 10 ATM nodes, five Nectar nodes, and the Field Reliability Test.

RAIDframe currently runs at the user and kernel levels. TIP-1 and TIP-2 currently run on Alphas. SPFS runs Scotch-3 version one which uses five Nectar nodes only.

Departures & Arrivals

Three members of the Parallel Data Lab have left its ranks: Jiawen Su, Mark Holland, and Rachad Youssef.

Jiawen left the group at the end of the fall semester. While a member of the Lab, he worked on asynchronous name resolution for the informed prefetching project. He's now working on the NetBill project at CMU.

Mark, a postdoctoral research faculty member of the Lab, left June 30 for the Peace Corps and Malawi (Africa). Mark finished and defended his dissertation "On-line Reconstructon In Redundant Disk Arrays" in 1994.

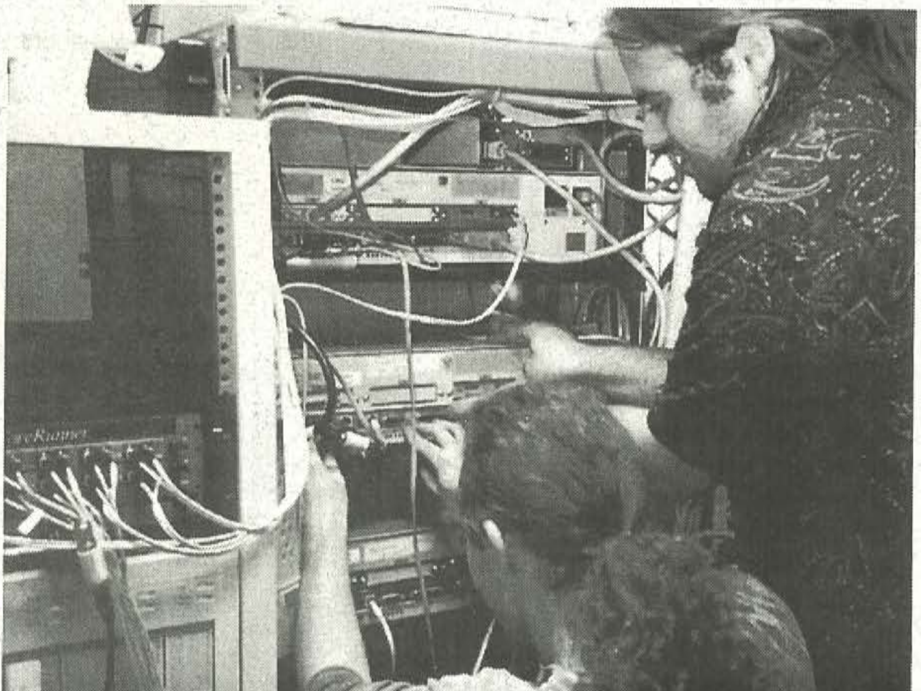
Rachad Youssef, a master's student in the Information Networking Institute, accepted a job at Oracle Corporation. For his master's thesis, Rachad investigated the feasibility of using disk arrays for mobile computers. At Oracle, Rachad coordinate the internatinalization of Oracle products.

Three member have joined the PDL. One student, Fay Chang, joined the Parallel Data Lab half-time last fall; two more, Erik Riedel and Khalil Amiri, begin graduate school this fall in the Electrical and Computer Engineering Department.

Fay came to the CS department at CMU from Duke University where she earned a B.S. in electrical engineering. She is currently working with research faculty member Steve Lucco on safe user networking services.

Erik Riedel has worked on the Geographic Environmental Modeling System (GEMS) at CMU and Models-3 at the U.S. Environmental Protection Agency. His work for the Lab will cover spatial data systems.

Khalil Amiri comes to the PDL via the University of Pennsylvania where he was a master's graduate student in computer science and engineering. At Penn, Khalil worked on design and implementation of distributed filesystems and transaction processing systems. He will likely start work on the new NASD project.



"Watch the fingers!"

Jim Zelenka *might* be warning Daniel Stodolsky. Jim is one of the PDL's staff programmers and Danner is a fifth-year graduate student; both spend most of their waking hours opening up the guts of hard drives and stringing cable (which usually requires removing floor tiles). Here, they're reinserting Lagavulin after upgrading it to OSF/1.3.2.

Jim and Danner spend a lot of time setting up and maintaining the working environment for all of the projects in the lab. They're involved in building the Scotch Parallel File System (SPFS) and supporting the informed prefetching and caching project.

PDL Web Site: Pages Are Reorganized and Updated

As with most Web sites, the Parallel Data Lab's site has been transformed over the past year as the group seeks to expand its on-line presence and distribute its products, including papers, presentations, and software, to other researchers and developers of parallel I/O systems.

Since the group had already taken Skibo Castle—formerly Andrew Carnegie's summer home—as a visual metaphor for its storage systems research, it was an ideal choice for adding a graphic face to the PDL Web site. For this reason, most of the contents are organized according to what room in a castle a visitor would find them. For example, research highlights and papers, group bibliographies, confrences, and on-line references are all found in the Library. Important links to academic, industry, and government sites are also available through the Library.

The other three rooms are the Lab, the Library, and the Salon. The Lab contains software and documentation released by the PDL as well as some Web publishing tools and experimental papers created using them. The Gallery, appropriately enough, presents pictures of Lab members and the Lab itself. These pictures are currently being updated to reflect the changes in the Lab over the past year. The Salon is a room for fun, non-research topics, such as the history of Skibo Castle.

The content that did not fit neatly into the room metaphor has been grouped into separate pages with links named Research World, Recent Results, and Lab Jobs.

The Research World page lists and links the numerous academic, industry, and government affiliates of the PDL. These include CMU-based and external research projects, industry collaborators, and funding organizations such as ARPA and NSF.

The Recent Results page is a short list of links to the group's Web-based ARPA reports. And the Lab Jobs page is where the group posts information about positions in the Lab, including post-doctoral, graduate, undergraduate, and staff positions. Because the group is also interested in keeping Web visitors up to date on what kinds of positions are available in industry, it plans to share informal job announcements occasionally.

All rooms and high-level pages are identified with icons and linked from the PDL front door page (also known as the home page). In order to accommodate different viewing needs and the time necessary for downloading graphics, a text index of the content is also available. Most of the links from this high-level index avoid the graphic room pages.

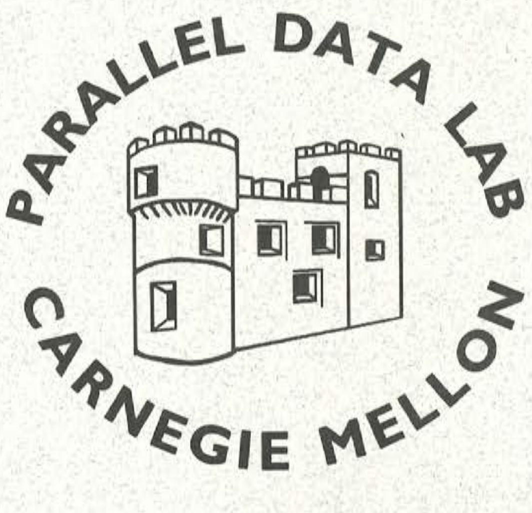
Work continues on adding HTML versions of group research papers and presentations; later this fall the *PDL Packet* will also be available. The group also plans to transform its growing database of papers on parallel I/O into a series of smaller, searchable HTML bibliographies.

Parallel Data Laboratory Shirts Available

Polo shirts bearing the new shirt logo of the Parallel Data Laboratory were ordered from Lands End. The logo, which was originally based on the illustration used in the newsletter nameplate, slowly evolved into a less-detailed section of Skibo Castle in order to accommodate the limitations of the embroidery technology.

Members of the Parallel Data Consortium, as well as all lab members, will be gifted with a polo shirt when they attend the 1995 fall workshop and retreat.

The logo shown at the right is exactly the same size as that on the shirts.


PARALLEL DATA LAB
CARNEGIE MELLON