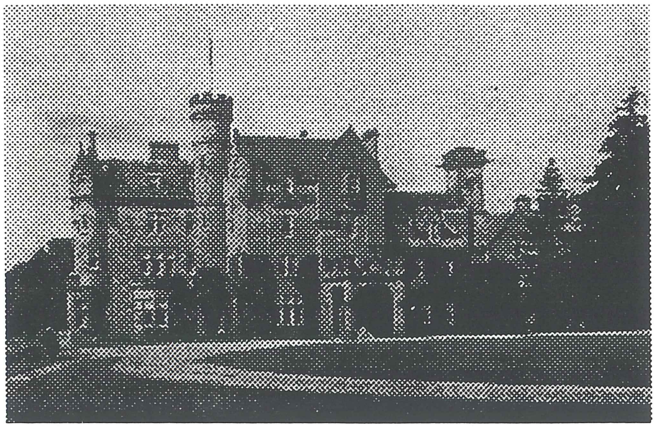


PARALLEL DATA LABORATORY

UPDATE

Skibo Castle, Scotland

August 1994

JUNK MAIL, NOT!

Welcome to the first issue of the Parallel Data Laboratory's UPDATE newsletter. In these letters we will be bringing you news, views, and attitudes. And if that isn't enough (and perhaps it shouldn't be) these newsletters will be accompanied by recently published papers and reports.

Editorial responsibility (that is, blame) falls on the Lab's leader, Professor Garth Gibson, (412) 268 5890.

SKIBO CASTLE

Andrew Carnegie's Summer Home

In honor of Carnegie Mellon University's great patron, Andrew, and the campus' recently departed Skibo Hall, the Parallel Data Laboratory has taken Skibo Castle as its logo. A fortress of storage perhaps; the central feature of a computing community, perhaps; an interesting motif from which we can draw a wide range of machine names, certainly. And, incidentally, our vision of the ideal retreat and workshop venue.

Kyle of Sutherland

Taking Scotland as the theme for selecting names in the PDL has proved effective and entertaining. For example, Skibo is in the Kyle of Sutherland. You might expect Sutherland to be in the south of Scotland, but instead it is on the northern shore of the Dornoch Firth (a strait), north of Inverness, and north of most of Scotland. But it is south of the Shetland and Orkney Islands that were a base for Viking invasions in the 9th century.

Scotch and its Sequels

The parallel storage system testbeds under construction in the PDL are known as Scotch generations. The first Scotch system (Scotch I for the Hollywood slaves among us) is composed of 10 AT&T-GIS 6299 disk arrays and 150 Seagate disks (ST12400 and ST31200). Scotch II, also under construction, is composed of a DEC StorageWorks SW800 and 60 HP2247 disks. With Star Wars as a model, perhaps there will someday be nine generations of Scotch parallel storage systems.

Bens

With so many arrays, controllers, cabinets, and shelves we feel compelled to have names better than "third from the top in the fourth cabinet from the left". Address tuples such as "host machine, SCSI adapter slot, target ID and LUN" would perhaps do if we don't move components much, but it just isn't fun. So we are going to name cabinets, arrays, and shelves (SCSI targets) after Scottish mountains like Ben Nevis, Ben Wyvis, Ben Vrackie, and Cairngorm.

Drams

Of course all this naming stuff originated with our need to give Internet names to PDL workstations. Here we exploit the wealth of Single Malt Scotch Whiskeys. For example, our initial machines have given names Macallan, Oban, Talisker, Laphroaig, Cockburns, Balvenie, Cardhu, Glenlivet, and Bowmore and family name DSSC.CS.CMU.EDU or PDL.CS.CMU.EDU.

PDL's FIRST Ph.D. GRADUATE

Dr. Mark C. Holland is the first Ph.D. graduate from the Parallel Data Laboratory. Mark's dissertation, "On-Line Data Reconstruction In Redundant Disk Arrays," is (in my humble opinion) the definitive "cookbook" resource for understanding the trade-off between performance, reliability, and cost in storage systems for on-line transaction process-

INDEX

ing (and for similar highly available systems). Mark is staying with the lab this year as a postdoc so that he can implement and evaluate his disk array architectures on the Scotch II system.

Mark's dissertation abstract.

> There exists a wide variety of applications in which data availability must be continuous, that is, where the system is never taken off-line and any interruption in the accessibility of stored data causes significant disruption in the service provided by the application. Examples include on-line transaction processing systems such as airline reservation systems, and automated teller networks in banking systems. In addition, there exist many applications for which a high degree of data availability is important, but continuous operation is not required. An example is a research and development environment, where access to a centrally-stored CAD system is often necessary to make progress on a design project. These applications and many others mandate both high performance and high availability from their storage subsystems.
>
> Parity-based redundant disk arrays are very attractive storage alternatives for these systems because they offer both low cost per megabyte and high data reliability. Unfortunately such systems exhibit poor availability characteristics; their performance is severely degraded in the presence of a disk failure. This dissertation addresses the design of parity-based redundant disk arrays that offer dramatically higher levels of performance in the presence of failure than systems comprising the current state of the art.
>
> We consider two primary aspects of the failure-recovery problem: the organization of the data and redundancy in the array, and the algorithm used to recover the lost data. We apply results from combinatorial theory to generate data and parity organizations that minimize performance degradation during failure recovery by evenly distributing all failure-induced workload over a larger-than-minimal collection of disks. We develop a reconstruction algorithm that is able to absorb for failure-recovery essentially all of the array's bandwidth that is not absorbed by the application process(es). Additionally, we develop a design for a redundant disk array targeted at extremely high availability through extremely fast failure recovery. This development also demonstrates the generality of the presented techniques.

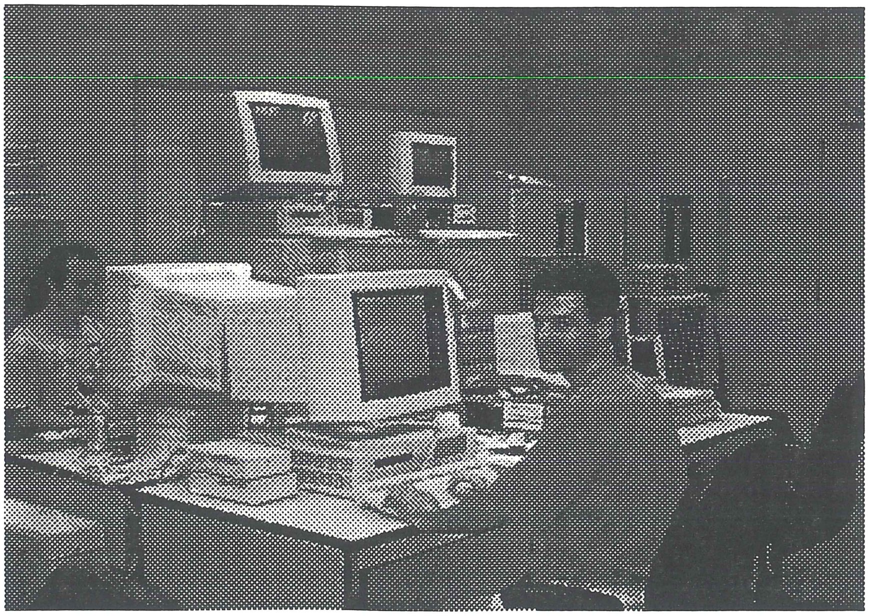
INCLUDED FOREST BY-PRODUCTS

"On-Line Data Reconstruction in Redundant Disk Arrays," Mark Holland, CMU-CS-94-164 technical report, May 1994.

"A Redundant Disk Array Architecture for Efficient Small Writes," Daniel Stodolsky, Mark Holland, William V. Courtright II, Garth A. Gibson, CMU-CS-94-170 technical report, July 1994. An abbreviated version of this will appear in ACM Transactions on Computer Systems, August 1994.

"Exposing I/O Concurrency with Informed Prefetching," R. Hugo Patterson, Garth A. Gibson, Proc. of Third Int. Conf. on Parallel and Distributed Information Systems, Austin TX, September 1994.

"RAID: High-Performance, Reliable Secondary Storage," Peter M. Chen, Edward K. Lee, Garth A. Gibson, Randy H. Katz, David A. Patterson, ACM Computing Surveys, v 26 n 2 June 1994, pp. 145-185.



SYSTEM SOFTWARE FOR MULTICOMPUTERS

Research in the Parallel Data Laboratory is now substantially supported by three organizations within CMU. Our initial and still growing source of support is the NSF-sponsored Data Storage Systems Center directed by Professor Mark Kryder. The PDL performs the majority of the research in the DSSC's thrust area for storage and computer systems integration.

In 1993, the PDL formed the Parallel Data Consortium for companies with little interest in the other DSSC thrust areas (magnetic recording, magneto-optic recording, and electronic subsystems). AT&T Global Information Systems (formerly NCR) and Data General seeded this consortium. Because PDC membership extends to all DSSC affiliate and associate members, companies such as IBM, Digital, Hewlett-Packard, Seagate, and Storage Technology have taken an active interest in the consortium.

Late in 1993 the lab took a third sponsor, the ARPA/HPCC-sponsored Systems Software for High-Performance Multicomputers project. Through this project, PDL file system research attains the resources to grow to a sustainable level of effort (answering a concern of the DSSC's 3rd-year NSF review committee).

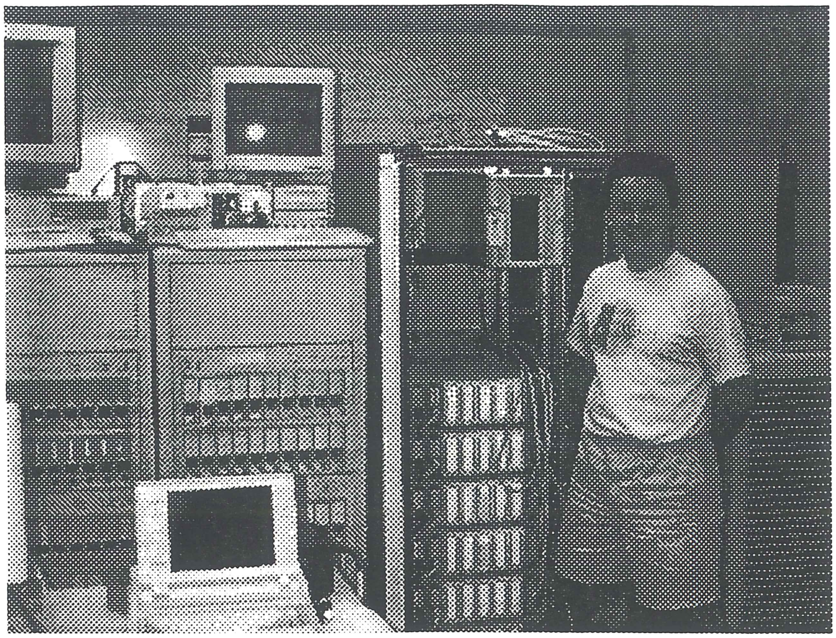
Effective Utilization of Heterogeneous Multicomputers

A multicomputer is a collection of individual processing elements connected by a high-speed network. While it is possible today to build a "tera-op" multicomputer, effective utilization of that machine remains a significant problem. This project is designing, implementing, and will use a cohesive operating system and runtime environment for the emerging generation of distributed multicomputer platforms.

Our approach to cost-effective high performance for HPCC applications is to harness the power of broadly heterogeneous multicomputers. By heterogeneous we mean a variety of processor architectures, an irregular and changing network topology, and a varying set of concurrent tasks with different and changing degrees of parallelism.

To demonstrate our approach we are constructing a new experimental software testbed for multicomputer systems. This testbed leverages hardware constructed for the ARPA Nectar Gigabit Network testbed and the NSF Data Storage Systems Center's Scotch parallel storage server. We are constructing

- a partitioned, safely-extensible, UNIX-based operating system (Bridge) with performance comparable to the best monolithic kernels,
- a prefetching file system (TIP) for large scientific datasets exploiting parallel servers and parallel paths in switched networks,
- an object-based programming environment (Dome) that transparently exploits changing resources for scientific applications,
- a distributed shared-memory programming environment (Midway) minimizing shared-data message traffic and avoiding page fault overhead with software write detection,
- operating system software and protocol support for high-bandwidth, low-latency networks,



- message system support for concurrent incremental checkpointing, and
- multicomputer monitoring and tracing facilities supporting parallel performance tuning and debugging.

This project's approach to system software for multicomputers owes much to its parent project, the Mach operating system, and in particular the Mach-US multi-server partitioning of system services, completed in 1994.

SCALABLE I/O INITIATIVE

The PDL will soon assume a small role in the about-to-be-funded Scalable I/O Initiative. This large project brings together a wide range of high-performance computer systems researchers under the leadership of CalTech's Professor Paul Messina. The Scalable I/O Initiative seeks to develop an integrated set of tools and software systems to provide a scalable I/O facility for HPCC grand challenge applications. The PDL's involvement is primarily focussed on exploring the utility of informed prefetching for maximizing storage efficiency for these grand challenge applications.

MOBILE ARRAYS

In cooperation with Hewlett-Packard and CMU's other NSF Engineering Research Center - the Engineering Design Research Center (EDRC) - PDL researchers have begun to look at the utility of small arrays of very small disks in mobile computers. The opportunity to spin up only one very small, lower-power disk when data must be transferred offers power saving potential without constraint on capacity. For mobile applications requiring substantial storage, anticipatable accesses, or poor wireless connectivity, such arrays may allow more compact and power efficient mobile computers.

1994 RETREAT PLANS

DSSC Annual Review: Oct 31 - Nov 2

The 1994 annual review of Data Storage Systems Center progress will be Monday October 31 through Wednesday November 2. This review brings DSSC member company representatives to CMU to evaluate DSSC research progress.

PDL Retreat and Workshop: Nov 2 - 4

The 1994 annual Parallel Data Laboratory retreat and workshop will be help

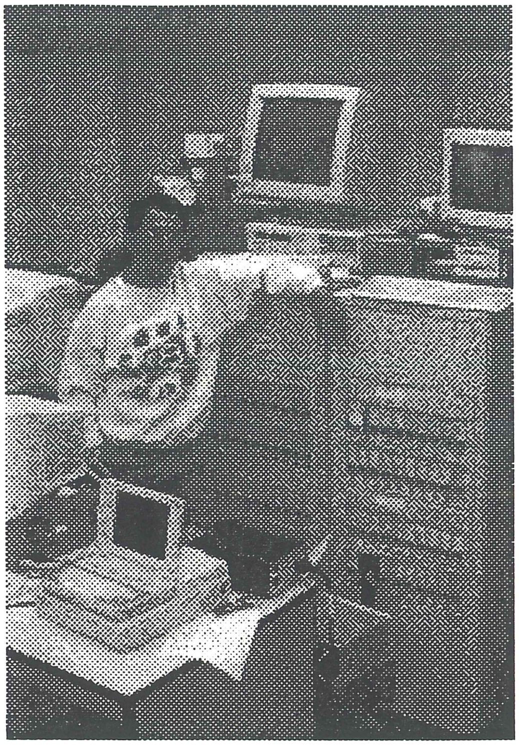
IMAGE INDEX

- Skibo Castle: the PDL motif
- Mark Holland at work on Scotch I (middle background) and Scotch II (right background) with Daniel Stodolsky in the left corner
- Garth Gibson proudly showing Scotch I and Scotch II storage hardware
- Bill Courtright, at home with the disk arrays built by his employer and Ph.D. sponsor: AT&T-GIS
- Rachad Youssef utilizing the AT&T-GIS 6298s to develop a workload generator for the disk reliability field test
- Daniel Stodolsky tending to the Scotch I equipment integration effort.

onWednesday November 2 through Friday November 4. This year we will extend our traditional emphasis on storage with a broad treatment of research collaborations underway in the Multicomputing Systems Software project. The following is a tentative list of talks:

- Scotch I disk reliability field test
- Scotch I parallel file system design
- Scotch II RAID device driver design
- Scotch II declustering evaluation
- Error handling in RAID software
- Disk arrays for mobile computers
- TIP evolution: asynchronous name resolution and smart cache management
- Parallel I/O interfacing to Dome parallel programs
- Software fault isolation: a tool for high-performance, safe systems software

The location of the retreat is not settled yet, but it will be more than one hour and less than three hours drive from Pittsburgh on a PDL chartered bus. We expect attendance of ten to fifteen guests and twenty to twenty-five CMU researchers. Bring warm, comfortable clothes for in addition to a generally casual atmosphere, the retreat schedule includes a two or three hour outdoor hike.



For more information on what it takes to get an invitation (Pennsylvania travel, Nov 2 & 3 accommodations and food paid for by the PDL) contact Garth Gibson at garth+@cs.cmu.edu or 412-268-5890. As a general rule we will try to invite one to three representatives of each of our member companies.

ACQUISITIONS DEPARTMENT

3607 Wean Hall

In January of 1994 the School of Computer Science came through with one of the scarcest of university resources - physical space. The PDL was given possession of 3607 Wean Hall, a 18'x24' room with a raised floor, 100 amps of 3-phase power, and machine-room cooling. Moreover, because of its proximity to the school's central machine rooms, disk array equipment in excess of space available in 3607 can be located across the hall and still be directly connected to 3607 machines by fast, wide, differential SCSI-2 cabling.

Scotch I Equipment Donations

Two equipment donations have provided the bulk of the equipment in Scotch I. First, Digital Equipment Corp., through their membership in the DSSC, donated five DEC 3000/400 (Alpha) workstations and ten fast, wide, differential SCSI-2 adapters. These machines are the servers in Scotch I. Second, Seagate Technology, also through their membership in the DSSC, donated eighty ST31200 and eighty ST12400 disk drives and the funds to purchase ten AT&T-GIS 6299 disk array subsystems. AT&T-GIS, in addition to furnishing a full scholarship for Bill Courtright's Ph.D. program, has provided substantial assistance with Scotch I and has donated two 6298 disk arrays (the predecessor of the 6299s).

Scotch II Equipment Donations

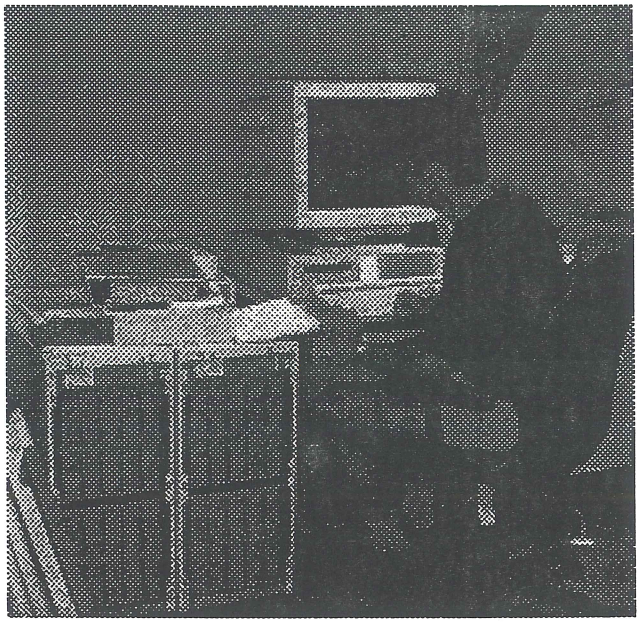
Two more equipment donations are at the heart of the Scotch II platform. First, the School of Computer Science purchased for and loaned to (long-term) us sixty-four HP 2247 disk drives. Second, DEC provided a StorageWorks SW800 and twelve drive-ready Storme shelves, once again through their membership in the DSSC. To this we have added a DEC 3000/500 (an Alpha with more Turbochannel slots) which will enable us to experiment with advanced disk array architectures by embedding such architectures in host-based array control software.

Mobile Arrays Equipment Donations

Hewlett-Packard is working with the PDL to equip our new mobile arrays research project. In transit is a donation of a series 700 workstation, which we will use to develop and experiment with arrays of small disks, and a dozen Omnibook mobile computers, with which we will experiment with capacity-intensive mobile applications. In progress is a donation of Kittyhawk 1.3" disks to form the core of this project's array of small disks.

WAYS TO HELP

- Join Parallel Data Consortium and attend PDL retreat Nov 2-4 1994.
- Involve more groups in your company: storage, operating systems, database, parallel programming, continuous media.
- Provide technology & marketplace vision and important application code or traces.
- Help us to discover more names of Scottish mountains or single malt whiskeys.



RS6000 Workstation Donation

IBM is also making a substantial contribution to the PDL equipment. We are in the process of arranging a donation of about nine RS6000 workstations. These machines, depending on final configurations and timing, may be used in Scotch I or Scotch II. We are also looking forward to ATM-based switched networks in which these workstations may play a central role.

The PDL is deeply grateful to DEC, Seagate, AT&T-GIS, HP, and IBM for this equipment.

PDL STAFF

Garth Gibson, who leads the PDL, is an assistant professor in CMU's School of Computer Science (SCS) and in its Electrical and Computer Engineering department (ECE).

Mark Holland has completed his ECE Ph.D. and is a postdoctoral fellow in the PDL implementing advanced RAID architectures on Scotch hardware.

R. Hugo Patterson II is a ECE Ph.D. candidate in the midst of dissertation work on an implementation and evaluation of Transparent Informed Prefetching.

Daniel Stodolsky, a SCS Ph.D. student, has completed his analysis of parity logging arrays and is beginning dissertation work on aggressive prefetching systems.

William V Courtright II, an ECE Ph.D. student, is about to propose that his dissertation research focus on disk array controller software structure for exception handling.

Jiawen Su is a new SCS Ph.D. student who has been working on asynchronous name resolution in the Transparent Informed Prefetching system.

Qingming Ma, a SCS Ph.D. student, is about to propose that his dissertation research bridge the gap between highly parallel network systems and highly parallel file systems for applications requiring parallel I/O streams.

Li-Kang Chen is a Math and Computer Science undergraduate student in the masters of software engineering program working on disk array management interface systems.

Rachad Youssef, a masters student in the Information Networking Institute (INI), is managing the Scotch I disk reliability field test and is pursuing a masters project on arrays of small disks for mobile computers.

Kin Chan is a Math and Computer Science undergraduate student working on information infrastructure interfaces and the PDL contribution to the Internet.

Alan Horn is an H & SS undergraduate student working on a database of PDL-related materials.

COMING SOON: VIDEO