A Scalable High-Performance I/O System*

Mark Henderson, Bill Nickless, Rick Stevens
Mathematics and Computer Science Division
Argonne National Laboratory
Argonne, IL 60439
{henderson,nickless, stevens}@mcs.anl.gov

Abstract

A significant weakness of many existing parallel supercomputers is their lack of high-performance parallel I/O. This weakness has prevented, in many cases, the full exploitation of the true potential of MPP systems. As part of a joint project with IBM, we have designed a parallel I/O system for an IBM SP system that can provide sustained I/O rates of greater than 160 MB/s from collections of compute nodes to archival disk and peak transfer rates that should exceed 400 MB/s from compute nodes to I/O servers. This testbed system will be used for a number of projects. First, it will provide a high-performance experimental I/O system for traditional computational science applications; second, it will be used as an I/O software and development environment for new parallel I/O algorithms and operating systems support; and third, it will be used as the foundation for a number of new projects designed to develop enabling technology for the National Information Infrastructure. This report describes the system under development at Argonne National Laboratory, provides some preliminary performance results, and outlines future experiments and directions.

1 Introduction

For a long time parallel I/O has been neglected as work in parallel computing systems has focused on the need to develop better communications networks, low-cost packaging, and microprocessors. Today, many parallel systems are on the market, and applications developers are busy porting and developing codes: it is now the case that parallel I/O can no longer be left on the back burner if massively parallel processing (MPP) systems are to become widespread replacements for mainframes and vector-based supercomputers. As increasingly large numbers of parallel systems are installed, it is important that the parallel I/O problem be addressed with the same energy and creativity as the parallel compute problem. This paper is an overview of a project at Argonne National Laboratory (ANL) to build a parallel I/O systems testbed and to develop parallel I/O software and applications that will test and use this software.

One principal goal in the I/O project is to create a testbed rich enough that a large number of experiments can be conducted without requiring new hardware to be developed and installed. This flexibility is important in the short term as we explore the possibilities for solving the I/O problem. Our overall goal in the project is to develop the framework for creating balanced high-performance computing systems in the future that are highly I/O capable. Another objective of the project is to provide the capability to test both parallel I/O systems and to validate those systems in the context of a robust mass storage environment.

The Argonne I/O system uses two layers of high-performance networking. There is a primary layer for interconnection between compute nodes of the IBM SP and a set of I/O servers. This primary layer uses the Fiber Channel Standard (FCS) for transport and connects 32 of 128 processor nodes to eight I/O high-performance servers. Five Ancor FCS switches provide the ability to route any I/O transfer to any I/O server processor. Server processors are IBM RS/6000 970Bs with 256 MB RAM and 6 GB of local disk. A secondary networking layer connects the eight I/O servers to four high-performance IBM 9570 RAID arrays (220 GB total) [Katz 89] and an automated tape library (DST-800 with three AMPEX DD-2 19-mm helical scan tape drives) with 6.4 TB of tape. The secondary networking layer is built on a NSC HIPPI

*This research supported in part by the Office of Scientific Computing, U.S. Department of Energy, under Contract W-31-109-Eng-38.

crossbar enabling any I/O server to access any RAID drive or tape unit. This I/O system has been designed to scale to systems with hundreds or thousands of processors. The primary and secondary networks can be expanded by adding additional crossbars into more complex networks if needed (see Fig. 1).

Software is being developed that allows user applications to migrate data to and from the tape subsystem, RAID arrays, local disk on the I/O servers, and the local disks on the compute nodes (each processor in the SP has a local 1-GB disk). Parallel I/O libraries are under development that enable user applications codes to manage hundreds of I/O streams and to checkpoint and restart with different numbers of processors. To manage the archiving of data, NSL-Unitree is run on the I/O server nodes and manages data on the RAIDs and tape subsystem. ANL is collaborating with the National Storage Laboratory (NSL) on the development of and testing of the successor to Unitree —the High-Performance Storage System (HPSS)— and with the Scalable I/O Initiative [Bers 93] to develop new parallel I/O concepts and implementations.

2 Applications Requirements

As the processing power of parallel supercomputers has increased over time, there has not been a corresponding increase in the I/O capability of those systems [Fren 91], [Redd 90]. Many existing parallel applications such as QCD and Monte Carlo problems do not have a large I/O requirements. However, many problems in computational chemistry, computational fluid dynamics, seismology, and emerging applications under development for the NII and other areas require high-performance I/O [Work 93]. This I/O is used for checkpointing large runs, visualization archival storage of run histories, and support of digital libraries. On current parallel supercomputing systems, applications require from hundreds of megabytes to hundreds of gigabytes per run and need I/O bandwidths in excess of 100 MB/s [Mill 93].

2.1 Global Climate Models

Global climate modeling is a good example to illustrate some of the issues. Estimates of performance and data requirements for parallel climate models indicate the scale of the I/O problem [Work 93]. In each case the dataset size is for the generation of the history "tape" of the run (i.e., the data that would normally be archived and stored for postprocessing).

- A 100-year, T42 run would generate 143 gigabytes
- A 100-year, T85 run would generate 1,144 gigabytes
- A 1000-year, T85 run would generate 20,000 gigabytes

In addition to archival data, periodic checkpointing is done that requires writing the contents of processor memory (in our case up to 16 GB) to secondary storage and then possibly reconfiguring that data for a restart that may involve a different number of CPUs. A typical working environment may need dozens of checkpoint/restart files available. The time to generate a checkpoint will determine the frequency of checkpointing. With the current I/O system we estimate completing a full memory checkpoint in under 3 minutes, thus enabling runs to proceed with little concern for restart. Archival datasets are often stored for years and involve hundreds of retrieval cycles. Analysis runs (which are less compute intensive than the simulation itself) typically complete in 1/10 or 1/100 the time of the data generation run; therefore, sustainable I/O rates would be the primary performance limitation. Analysis runs will be the main drivers for I/O resources.

2.2 Video Server Applications

In a related project —the Multimedia Supercomputing Project (MMSuper)— we are exploring the close coupling of multimedia technology (images, sound and text) and high-performance computing architectures and applications. By enabling systems such as the IBM SP to process voice and image data along with traditional scientific data resulting from computations, we hope not only to expand the types of applications that are suitable for parallel computing, (e.g., digital libraries) but also to facilitate the use of modern interface technology for those using the SP system for scientific applications (e.g., coupling animation and voice annotation with scientific data sets). The MMSuper project will develop and explore the features and architectures needed to support digital libraries and interactive multimedia use of parallel supercomputers.

Current parallel supercomputer systems like the IBM SP are well suited for "traditional" scientific computing (i.e., rapidly evolving a simulation described by a set of numerical partial differential equations forward in time and periodically capturing the state of the system). However, future uses of supercomputers will include not only numerical computations but also

the reduction of output data to meaningful statements about reality. Today one would like to use the supercomputer as a general-purpose tool for computational science including the use of the system for supporting visualization, collaborative development of systems, collaborative exploration and analysis of both experimental and computational data and the storage and archive of such data and collaborations. To enable these uses of MPPs, additional hardware and software support is needed.

By combining parallel disk, high-speed network, video compression hardware and software, and digital signal-processing hardware, a parallel supercomputer with a high-speed disk array and automated tape library would be an ideal interactive/video on demand server. Prerecorded video would be stored on tape and on the disk array in a compressed form (the exact compression required depends on performance and hardware considerations) within a video database that can support retrieval and replication (for added I/O bandwidth when needed) of video sequences. When a video request is received, a processor is assigned to find the sequence (via simple text indexing in the case of movies and existing program materials), begin decoding (compression translation if needed say from MPEG to H.261), and route to the appropriate output port (typically ATM network or composite video out). Since the video sequence is stored on a random access device (in this case a disk array), multiple reads can be under way with different offsets corresponding to different retrieve request times (or time shifts). If more simultaneous read requests are received than the disk system can handle, additional copies of the file are created from the tape system (see above) so that in a few minutes a complete copy of the file can be spawned on additional disks to provide better performance. Better transfer rates can be achieved if the file is buffered to RAM before decoding and transmitting.

Disk I/O bandwidth required for video server applications can be estimated as follows: If we start with a 128-processor system and wish to generate 8-K video streams at 25:1 compression using something like ATM with 155 Mb/s service, then each processor needs to handle 65 video sessions. If we assume each processor has 2 GB RAM, providing buffer space of 24 MB for each stream or (84 seconds of each stream), then the total disk-to-processor I/O bandwidth needed all 128 nodes is 2.4 GB/s, which is just possible today on a SP1 with a FCS card in each node.

2.3 Other Applications

Other applications that will be used to test the system include parallel computational chemistry, molecular dynamics, astrophysical fluid flows, genomic databases, intelligent highway vehicle systems models, and structural mechanics. The output from these models will be used to develop animated visualization sequences that require up to four megabytes of data per frame. Thus a five-minute high-definition video could require 36 gigabytes of storage and playback rates as high as 120 MB/s. The source data for such animations may require several orders of magnitude more data to transferred and processed, thus requiring substantially higher I/O rates. higher.

3 I/O System Implementation

Figure 1 is a block diagram of the full Argonne SP installation, showing the scale of the installation and the architecture of the interconnections. The Argonne SP installation contains 128 compute nodes based on IBM RS/6000 workstation components. Each node contains 128 megabytes of RAM and a one gigabyte directly attached disk. Every node is connected to the high-performance switch via a microchannel adapter. With the current switch adapters (TB0), a node can achieve a maximum data rate of approximately 6 megabytes/second through the high-performance switch. (The next-generation adapters, which will be available Q2 1994, will increase node-switch bandwidth to about 30 MB/s.) Currently there is no overlap between compute processing and data transfer over the high-performance switch.

In addition to the high-performance switch adapters, 32 of the nodes have Ancor Fibre Channel adapter cards. We have measured using the channel interface a data rate of at least 19 megabytes/second from a compute/IO node through the Fibre Channel network to the rest of the I/O system. Unlike the high-performance switch adapters, the Fibre Channel adapters should allow some overlap between processing and data transfer. The exact degree of overlap will be determined in this project.

For the purpose of an unambiguous nomenclature, we have designated the 96 nodes without a Fibre Channel adapter as compute nodes, and the 32 with Fibre Channel adapters as compute/IO nodes. Compute/IO nodes are completely usable as compute nodes, and completely compute-bound applications will suffer no penalty from using the 32 compute/IO nodes for computation.

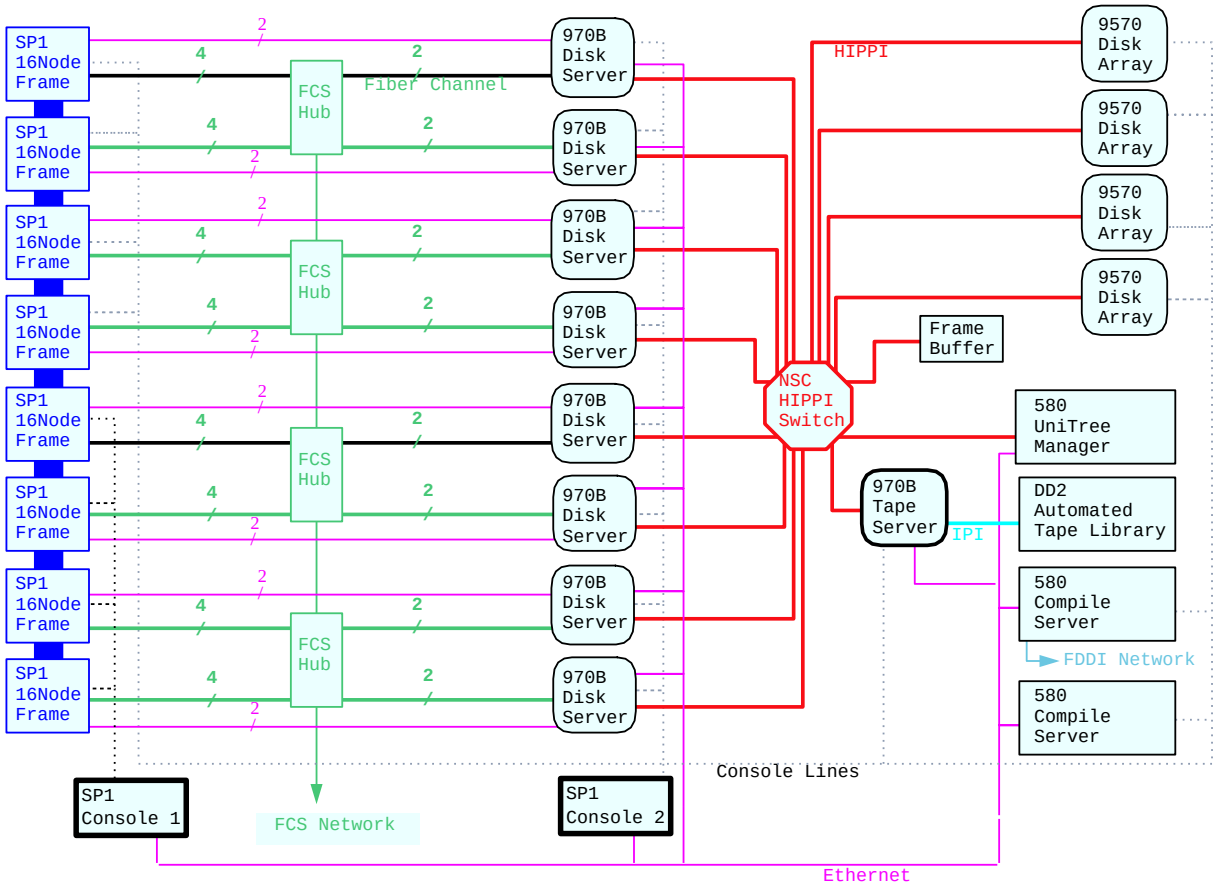

Figure 1. ANL's 128-node POWERparallel System

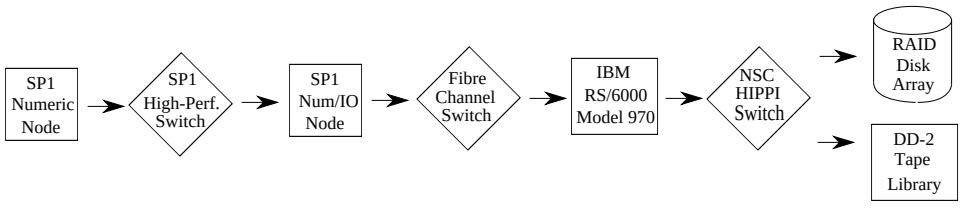

Figure 2. Logical Diagram of the Argonne I/O System

Throughout the design of the I/O system we have provided support for application–level multiplexing and demultiplexing of data. The first example of this principle can be seen in the three-to-one ratio between compute nodes and compute/IO nodes. I/O intensive applications will be able to intelligently spread the data among the compute nodes, as it will be delivered from the I/O servers via a relatively large number (up to 32) compute/IO nodes. If the application or some supporting library is structured to handle the data distribution internally, then the I/O can be seen as simply a class of message to be passed via the high-performance (message passing) switch. As the compute/IO nodes are complete computers in their own right, they are capable of running arbitrarily complex heuristics for scheduling I/O transactions efficiently. [Dura 93]

There is also support for TCP/IP over Fibre Channel and the high-performance switch, so data access can be transparent, yet relatively high-performance, for applications not taking full advantage of the capabilities of the system.

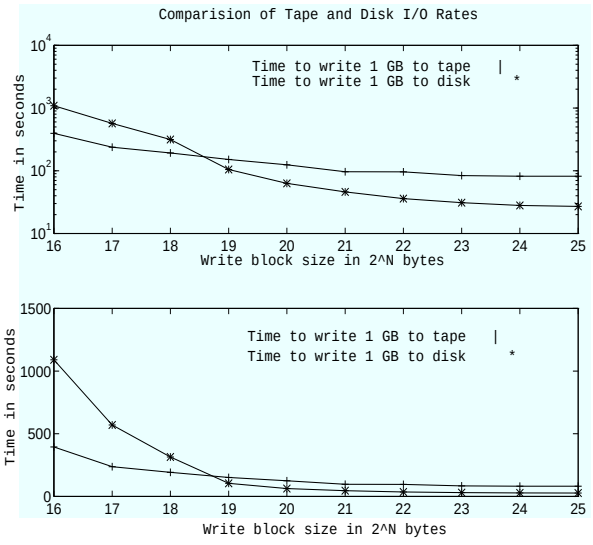

Figure 3. Tape vs. I/O Rates

The 32 Fibre Channel adapters in the compute/IO nodes are connected to five Ancor Fibre Channel switches, which are in turn connected to Fibre Channel adapters (2 each) in eight IBM RS/6000 model 970B computers. These RS/6000-970B computers are equipped with dual microchannel buses allowing independent input and output. Each RS/6000-970B also contains 256 megabytes of RAM, and six gigabytes of directly attached SCSI disk. Each Fibre Channel adapter will provide a data rate of 10-25 megabytes per second, for an aggregate rate of 20-50 megabytes per second per RS/6000-970B. This peak rate is balanced with the corresponding SP capabilities, as each compute/IO node can accept only 5-20 megabytes per second from the high-performance switch for sustained I/O (e.g., 8 RS/6000-970Bs at 20 MB/S = 160 MB/S, which equals 32 Compute/IO nodes at 5 MB/S = 160 MB/S).

The RS/6000-970B computers (I/O servers) also contain HIPPI adapters, which are capable of transferring data to and from the RAID disk arrays. The 9570 RAID disk arrays are currently limited to a read/write transfer rate of 40 megabytes/second. The HIPPI–attached peripherals such as the RAID disk arrays and the frame buffer are not directly flexibly programmable (they appear to the I/O server as a device, not a computer), but they are controlled from the general-purpose I/O servers. They provide another opportunity for arbitrarily complex scheduling heuristics to be tested [Dura 93], [Rosa 93].

4 Scalability Issues

Figure 2 is a reduced logical diagram of the I/O system of the Argonne I/O system showing the relationship of the various components. From left to right there are fewer instances of each component, the data rates increase, and the optimal size of each I/O operation increases as well. This logical diagram of the system is the basis for a number of modeling efforts under way aimed at developing I/O scheduling algorithms and implementation strategies. At each layer in the network there are opportunities for variations in buffering, load balancing, I/O block merging/splitting, and changes in protocol.

Throughout the design of the I/O system we have attempted to eliminate bottlenecks and "hot spots." This has been done by creating a hierarchy of data movers, with several stages of complete interconnect between the layers. This flexibility is achieved by using general-purpose computers in the I/O structure with enough memory to provide plenty of buffer space for reorganization of data for efficient transfer at to the next level.

As applications start using the system, we will be able to tune the characteristics of the data flow. The parallel I/O software under development has many adjustable parameters. The best settings for these parameters will be determined by extensive I/O application benchmarking and analysis. For example, an

application might be I/O bound if it has only one Fibre Channel connection to the rest of the I/O system; however, 32 Fibre Channel connections may be too many.

One focus of our I/O research is to study the effects of a particular architectural change: removing the RS/6000-970B computers as a boundary between the Fibre Channel network and the peripherals themselves. How necessary is this secondary layer of multiplexing and buffering? With appropriately tuned applications, can it be eliminated? Answers to these questions will affect future I/O system architectures.

5 Component Experiments

As a first step in developing comprehensive I/O models and designing experiments for scheduling and buffering we are collecting comprehensive baseline data on individual components of the I/O system. Detailed analysis and modeling will appear in a future paper.

In the present configuration, disk I/O to a IBM 9570 RAID unit can be sustained at between several megabytes per second to over 35 MB/s depending on the block size used in the transfer (see Fig. 4). In Fig. 3 we compare the I/O speeds of a single RAID unit and one DD-2 tape drive for a variety of block sizes, it is clear from this figure that substantial disk or memory buffering is needed to provide a user application optimum tape performance. However, for relatively small transfers, the DD-2 tape system is actually faster than the disk array. These data are for raw I/O.

We currently are investigating the effectiveness of interleaving many small transfers from the I/O servers as a way of improving the effective bandwidth for small block sizes. We feel that this may be crucial to efficient MPP I/O as MPPs are by nature-distributed memory machines. Unlike traditional supercomputers, there is no single contiguous memory space that can be transferred to storage and network systems in a very few I/O operations. Current high-performance storage and network technologies assume that there is a large, fast, contiguous memory that can handle transfers to and from the machine in a very few large transfers. This allows the latency often associated with high bandwidth to be amortized over the long transfers of contiguous memory. The I/O servers in the ANL configuration are intended to provide buffer space analagous to the large, fast, contiguous main memory stores of traditional supercomputers.

5.1 RAID Disk Experiments

The IBM 9570 RAID disk arrays are organized internally to have multiple logical volumes, allowing transfers to logically independent elements of the RAID. In Fig. 4 we have measured the effect of multiple I/O servers writing to the same RAID array. An important observation is that the I/O performance is dominated by block size until block sizes approach 2 MB. In Fig. 4 the top line is for two simultaneous writes, and the bottom line is for one write.

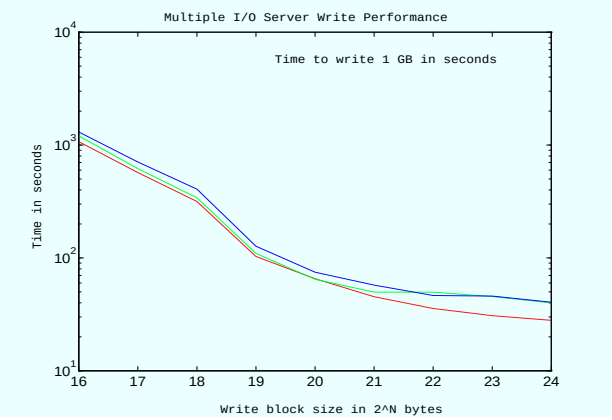

Figure 4. Effect of Multiple I/O Servers Writing to Same RAID Array

5.2 Tape System Experiments

Each DD-2 tape drive is connected to the tape server (a dedicated IBM 970) via an IPI interface; the tape server is HIPPI connected to the I/O servers. The Ampex DST-800 system has three drives installed and accessible from the 970. Current versions of Unitree do not stripe data across the multiple drives and an important I/O issue under investigation is to explore the feasibility of tape striping [Sale 86]. We are concerned with the raw performance of writing to all three drives simultaneously using both direct access and under Unitree control. Current experiments are able to achieve greater than 14 MB/s sustained per drive for two simultaneous transfers, giving an aggregate transfer rate of 28 MB/s (see Fig. 5); however, we have not been able to sustain aggregate transfer rates of 30 MB/s or more to three tape drives.

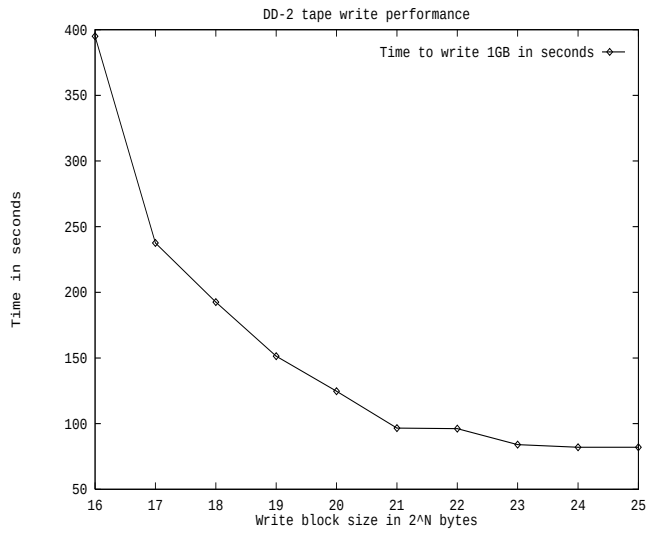

Figure 5. DD-2 Tape Write Performance

5.3 Ongoing Experiments

We are currently designing and conducting experiments to determine the answers to the following questions:

- What is the appropriate buffering strategy for a multilevel I/O fabric (i.e., do we buffer on the I/O nodes and I/O server or just on the I/O server)?
- What are the appropriate buffering strategies for isochronous I/O (e.g., video and audio streams)?
- Can we use IP protocols for the intermediate network layers and still have adequate performance end to end?
- What is the impact of interleaving reads/writes from the I/O nodes to the I/O servers?
- How much and what kind of I/O scheduling information or hints must be communicated from the I/O nodes to the I/O servers to ensure high performance?

Other experiments are under way to evaluate the use of FCS as the secondary network layer and to compare it with ATM [Lesl 93]. It appears that one of the most important feature of any intermediate I/O network is the degree of host CPU load required to sustain peak transfer rates.

6 Future Plans

We will be working with NSL-Unitree, HPSS, and the Vesta [Vest 93] (IBM PFS) parallel filesystem to perform early testing and to provide feedback into the development process. Our collaborators in the Scalable I/O Initiative [Bers 93], [Rosa 93] will be using the Argonne system as one testbed for their research. We hope the flexibility of the system will allow these and other researchers the opportunity to test techniques for efficient, parallel I/O and data storage in a production–scale facility.

Acknowledgments

We acknowledge the Department of Energy's Office of Scientific Computing and the National Aeronautics and Space Administration HPCC program for their support of the High-Performance Computing Research Facility at ANL. We also thank IBM for their generous support of both the facility at ANL and joint development activities which have made this project possible.

References

[Bers 93] B. Bershad et al. *The Scalable I/O Initiative*, white paper available from Argonne National Laboratory, 1993.

[Dura 93] Dannie Durand and Ravi Jain. *Distributed Scheduling Algorithms to Improve the Performance of Parallel Data Transfers*, Bellcore Technical Report, 1993.

[Fost 91] Ian Foster, Mark Henderson, Rick Stevens. *Proceedings of the Workshop on Data Systems for Parallel Climate Models.* Argonne National Laboratory, Mathematics and Computer Science Division Technical Memorandum ANL/MCS-TM-169, 1991.

[Fren 91] J. C. French, T. W. Pratt, and M. Das. *Performance of a Parallel Input/Output System for the Intel iPSC/860 Hypercube*, in SIGMETRICS Conf. Measurement and Modeling of Comput. Syst., pp. 178–187, May 1991.

[Katz 89] R. H. Katz, G. A. Gibson, and D. A. Patterson. *Disk System Architectures for High Performance Computing*, Proc. IEEE 77(12), pp. 1842–1858, Dec. 1989.

[Lesl 93] Ian M. Leslie, Derek R. McAuley, and David L. Tennenhouse. *ATM everywhere?*, IEEE Network, pp. 40–46, March 1993.

[Mich 93] John G. Michalakes. private communication, Argonne National Laboratory, 1993.

[Mill 93] Ethan L. Miller. *Input/Output Behavior of Supercomputing Applications*, UCB report 91/616, January 1991, University of California, Berkeley, CA, 1991.

[Redd 90] A. L. Narasimha Reddy and Prithviraj Banerjee, *A Study of I/O Behavior of Perfect Benchmarks on a Multiprocessor*, Proc. IEEE 17th Annual Intl. Symp. on Comp. Arch., May 1990.

[Rosa 93] Juan Miguel del Rosario and Alok Choudhary, *High Performance I/O for Parallel Computers: Problems and Prospects*, preprint Syracuse University, Northeast Parallel Architectures Center, Syracuse University, NY 13244-4100, 1993.

[Sale 86] K. Salem and H. Garcia-Molina, *Disk Striping*, in Proc. Intl. Conf. Data Engineering, pp. 336–342, 1986.

[Vest 93] Peter F. Corbett, Sandra Johnson Baylor, and Dror G. Feitelson. *Overview of the Vesta Parallel File System*, Research Report, IBM T. J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598.

[Work 93] R. Stevens (Editor). *Workshop on Grand Challenges Applications and Software Technology*, Pittsburgh, May 1993, available on `http://www.mcs.anl.gov/workshop.html`.