Expanding the Potential for Disk-Directed I/O

David Kotz

Technical Report PCS-TR95–254
Department of Computer Science
Dartmouth College
Hanover, NH 03755-3510
dfk@cs.dartmouth.edu

March 17, 1995

Abstract

As parallel computers are increasingly used to run scientific applications with large data sets, and as processor speeds continue to increase, it becomes more important to provide fast, effective parallel file systems for data storage and for temporary files. In an earlier work we demonstrated that a technique we call disk-directed I/O has the potential to provide consistent high performance for large, collective, structured I/O requests. In this paper we expand on this potential by demonstrating the ability of a disk-directed I/O system to read irregular subsets of data from a file, and to filter and distribute incoming data according to data-dependent functions.

1 Introduction

Despite dramatic improvements in processor technology, parallel-computer architecture, parallel languages and compilers, programming environments, and parallel algorithms, many programmers of scientific applications for massively parallel processors discover that their application's performance is limited by the rudimentary data-storage systems available on today's multiprocessors. When they find a multiprocessor that is configured with sufficient parallel-I/O hardware (unfortunately, many are not) they often discover that the file system software is not designed to meet their needs [CK93, KN94, PEK+95], or has poor performance [Nit92, KR94, FBD94].

As a result, there are several proposals for new interfaces, run-time libraries, compilers, languages, and file systems to support parallel applications on parallel computers. The focus of this paper is on a file-system technique called disk-directed I/O, which can dramatically improve the performance of reading and writing a large, regular data structure (like a matrix) between memory

This research was funded by NSF under grant number CCR-9404919, and by NASA Ames under agreement number NCC 2-849. The work was performed while the author was on sabbatical at Syracuse University.

Copyright 1995 by the authors

that is distributed across many processors and a file that is distributed across many disks [Kot94]. We explore the potential for disk-directed I/O in three other situations: data-dependent distribution (Section 3), data-dependent filtering (Section 4), and working with irregular subsets (Section 5). We present conclusions and a look toward the future in Section 6.

2 Background

There are many different parallel file systems [DSE88, Pie89, Dib90, DdR92, Mas92, FPD93, LIN+93, Roy93, CF94, Kri94, MS94, HdC95, HER+95]. Most are based on a fairly traditional Unix-like interface, in which individual processes make a request to the file system for each piece of the file they read or write. Increasingly common, however, are specialized interfaces to support multidimensional matrices [GL91, Mas92, BdC93, CFPB93, GGL93, BBS+94, SW94], and interfaces that support *collective* I/O [Mas92, BdC93, GGL93, BBS+94, CFH+95]. With a collective-I/O interface, all processes make a single joint request to the file system, rather than numerous independent requests.

In this paper we assume that the multiprocessor has an architecture like that in Figure 1, in which there are two types of processor nodes: those without disks, which are called compute processors (CPs), and those with disks, which are dedicated to the file system and which are called I/O processors (IOPs). Most, though not all, of the above parallel file systems are designed for machines with an architecture of this type.

There are several database machines that can filter tuples from blocks of data as they are read off disk, forwarding only those of interest on to the computational nodes (e.g., Tandem NonStop [EGKS90]). The Super Database Computer [KHH+92] has disk controllers that continuously produce *tasks* from the input data set, which are consumed and processed by CPs as they become available. Thus they have a load-dependent data-distribution mechanism. The Bridge (PIFS) file system [DSE88] tried to distribute data among memories to improve access locality.

The PASSION library can read submatrices that can be represented as a large contiguous region with some "holes" of unwanted data, by reading the full region of data and then "sieving" out the undesired data [TBC+94]. This sieve is not data-dependent, and is used to allow the library to make larger, more efficient requests to the file system.

Disk-directed I/O. Disk-directed I/O is a technique for optimizing data transfer given a high-level, collective interface [Kot94]. In this scheme, the complete high-level, collective request is passed

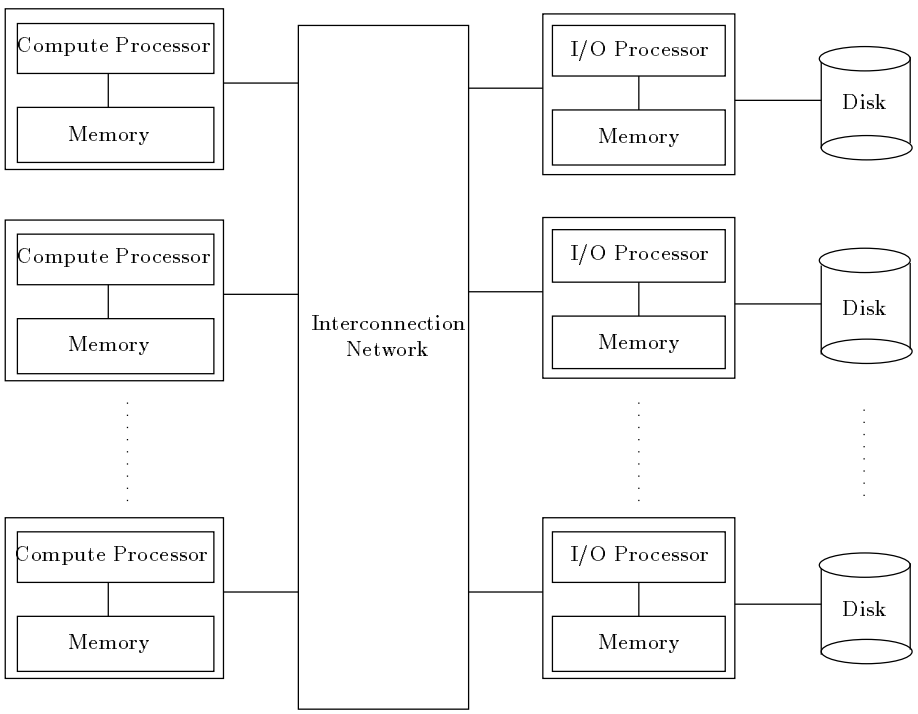


Figure 1: An architecture with two types of processor nodes: compute processors and I/O processors. The latter have attached disks and are dedicated to file-system service.

to the I/O processors, which examine the request, make a list of disk blocks to be transferred, sort the list, and then use double-buffering and special remote-memory "get" and "put" messages to pipeline the transfer of data between compute-processor memories and the disks. Compared to a traditional system with caches at the I/O processors, this strategy optimizes the disk accesses, uses less memory (no cache at the I/O processors), and has less CPU and message-passing overhead. In experiments with reading and writing one- and two-dimensional matrices, disk-directed I/O was as much as 18 times faster than traditional caching in some access patterns, and was never slower [Kot94].

An interesting application. Karpovich et al. [KFG94] describe the problem of storing and retrieving radio-astronomy data sets. The read-mostly data set is large and multi-dimensional: each data point represents an astronomical reading at some *time* at some *frequency* on some *instrument* pointed at some *region of the sky*. Needless to say the data set is extremely sparse. They store the data set by partitioning into buckets along a few of the dimensions, and sorting within a bucket along other dimensions. Applications rarely read the entire data set; instead, they request a subset of the data by specifying ranges for the time, frequency, and region. Using an index of buckets, only the necessary buckets must be read into memory. The buckets are then filtered to extract the items of interest, and (in a parallel application) distributed among memories of the multiprocessor according to the application's needs.

Clearly this application has very different I/O needs from those imagined for the disk-directed-I/O system in [Kot94]. It reads an irregular, discontiguous subset of data from the file. It filters out and discards some of the data it reads, after examining the data. Finally, it distributes the data among the memories in a data-dependent manner. In the remainder of the paper, we show how the concept of disk-directed I/O can also include these unusual requirements.

3 Data-dependent distributions

In the disk-directed I/O system described in [Kot94], matrices could be read from the file, distributing records among the memories in many different distributions. As each block was read from disk (in whatever order was convenient for the disk), the records within that block were sent to the appropriate location in the appropriate memory, based on the distribution function. In [Kot94] the distribution function was independent of the data; of course, a different, data-dependent distribution function could easily be used for the same purpose.

A traditional file system, however, is quite different. With a data-independent distribution, each processor independently computes the locations of the records it requires from the file, and reads those records. With a data-dependent distribution, however, there is no way for processors to request their own set of data. A reasonable solution is similar to two-phase I/O [dBC93]: each processor reads some convenient subset of data from the file, examines each record to compute the distribution function, and then sends the data to the appropriate processor.

In both cases we assume that the distribution function can only decide to which processor each record belongs, and send the record to that processor. Once there, the processor appends the record to its buffer for later processing.

3.1 Experiments

To gauge the impact of data-dependent distribution on performance, we devised an experiment to compare our disk-directed I/O and traditional-caching file systems. Of course, for the purpose of this experiment it matters little what distribution function we actually use — even a data-*independent* function would do. We used a cyclic distribution (`rc` in [Kot94]). Thus, the disk-directed system needed no change for this experiment. In the traditional caching system, the compute processors each looped reading blocks from the file, and for each record within each block, sent the record on to the appropriate destination processor. Logically, it made no difference which blocks were read by which processor, since most records would be redistributed anyway. For best performance on contiguous layouts [Kot94], we chose to have compute processors read the *blocks* in a cyclic distribution.

We ran these experiments on our simulator from [Kot94] called STARFISH. STARFISH ran on top of the Proteus parallel-architecture simulator [BDCW91], which in turn ran on a DEC-5000 workstation. We configured Proteus as in [Kot94], as shown in Table 1. In all cases a 10 Mbyte file was striped across disks, block by block, using one of two layouts within each disk: contiguous or random. Blocks were 8 Kbytes.

We repeated each experiment in this paper five times and report the mean value of each measure here. The largest coefficient of variation of any data point was 2.6%, so the trials were extremely consistent.

3.2 Results

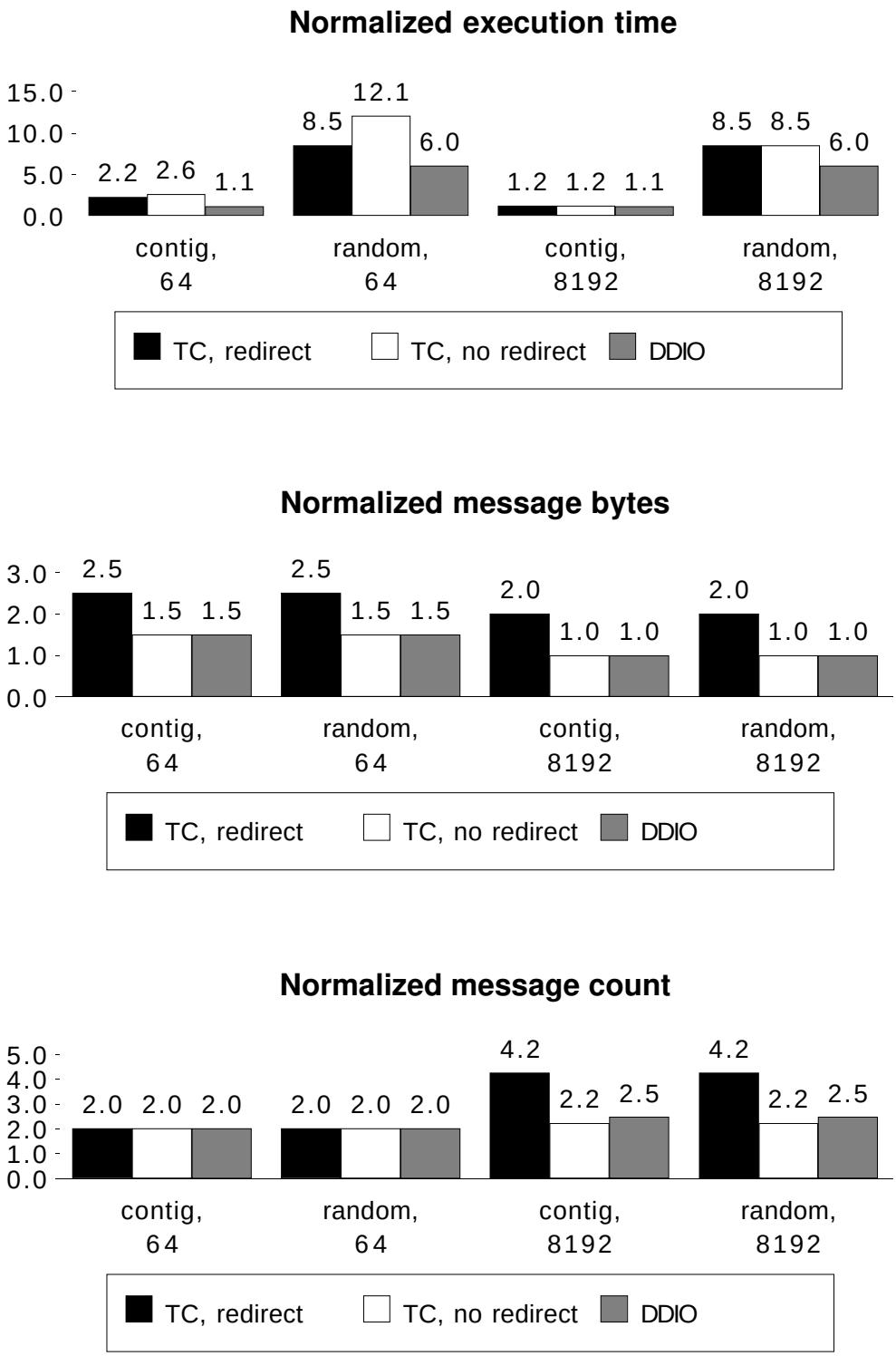


Figure 2: Results of data-dependent distribution experiments. Each graph is normalized against the "best" value for that measure. TC is traditional caching, and DDIO is disk-directed I/O. "contig" is a contiguous disk layout, and "random" is a random disk layout. Both 64- and 8192-byte records were used.

Table 1: Parameters for simulator.

MIMD, distributed-memory	32 processors
Compute processors (CPs)	16
I/O processors (IOPs)	16
CPU speed, type	50 MHz, RISC
Disks	16
Disk type	HP 97560
Disk capacity	1.3 GB
Disk peak transfer rate	2.34 Mbytes/s
File-system block size	8 KB
I/O buses (one per IOP)	16
I/O bus type	SCSI
I/O bus peak bandwidth	10 Mbytes/s
Interconnect topology	6×6 torus
Interconnect bandwidth	200×10^6 bytes/s bidirectional
Interconnect latency	20 ns per router
Routing	wormhole

Figure 2 shows the results of these experiments. These charts plot the execution time for the experiment, the number of bytes sent through the interconnect as messages, and the number of messages, each *normalized* against the best possible value for that measure. In all cases a smaller number is better, with 1.0 being the best. For execution time, the best possible value is computed from the amount of data read off disk and the peak throughput of the disk drives. Clearly, that execution time was only possible with no overhead and a contiguous layout. For message bytes, the best possible value is obtained when only the data itself is sent from the I/O nodes directly to the appropriate compute nodes. For message count, the best possible value is the number of records, since in our system each record is sent to its destination as a separate message. STARFISH often uses protocols with an acknowledgement sent for each message, so the normalized message count is usually greater than 2.

Each chart compares two disk layouts (contiguous and random), two record sizes (64 bytes and 8192 bytes), and two file systems (traditional caching (TC) and disk-directed I/O (DDIO)). For comparison, we add "TC, no redirect", in which each processor directly requested the data it needed; this would be impossible with a truly data-dependent distribution, but provides an interesting comparison.

Consider first the 8192-byte records. On the contiguous layout, all three systems performed

similarly (which is why we chose this distribution pattern), but using traditional caching to fetch and then redirect the data pushed all of the data through the network twice. The number of messages exploded as each block needed four messages: I/O request, I/O reply, send to another node, and acknowledgement from the other node. On the random layout, disk-directed I/O was faster because it could schedule the I/Os for better disk performance. Disk-directed I/O used slightly more messages here; these were startup overhead and would be negligible with a larger number of records, as can be seen in the bars for 64-byte records.

Consider the 64-byte records, which present an interesting picture. Despite nearly doubling the amount of message traffic, traditional caching with data-dependent redistribution was *faster* than the direct-access version. Here we were essentially using a pipelined form of two-phase I/O [dBC93]. Compute processors requested whole blocks from the I/O nodes, rather than small records. The larger requests reduced the overhead bottleneck experienced at I/O nodes, which in this case more than offset the extra work at the compute nodes and in the network. Disk-directed I/O was always better, however, because it could not only avoid the extra messages, but had less overhead and could optimize the disk traffic.

In our experiments the network was fast enough to not be a bottleneck. Thus, redistributing the data from the compute nodes was not a performance problem. Of course, systems with slower networks, or with networks being shared by other applications, may find disk-directed I/O to be even more valuable for its ability to avoid doubling the network traffic. An important example is a workstation cluster being used as a parallel computer.

4 Data-dependent filtering

If the disk-directed file system can make distribution decisions based on the data in the record, it is easy to see how it could filter out records according to some function of the data in the record. (Again, we assume that we do not send the data to a specific location within the destination processor; instead, the destination appends newly arriving records to its current buffer.) The same amount of I/O is needed, but there may be substantial savings on network traffic.

In a traditional system, each compute processor reads its data in chunks, filtering and compacting each chunk before reading more from the file system. The same amount of I/O is needed, and the same amount of network traffic is needed, as without filtering. Thus, traditional caching is essentially unaffected by filtering.

4.1 Experiments

We again used the cyclic (`rc`) distribution pattern. We filtered out either 0% or 90% of the records, according to a random filtration function (since again it does not matter specifically what function we use). Other system parameters are the same.

4.2 Results

Figure 3 shows the results of these experiments with two disk layouts (contiguous and random), two record sizes (64 bytes and 8192 bytes), two file systems (traditional caching and disk-directed I/O), and two filtration ratios (0% and 90%). Since we used a trivial filtration function, traditional caching has the same performance regardless of the filtration ratio. On random layouts disk-directed I/O can optimize the disk schedule, and on small records it has lower overhead, so it had better performance in those cases than did traditional caching. This chart shows that adding filtration does not affect performance, which is no surprise, but that it dramatically reduces the network traffic. Thus, filtering at the I/O processor rather than at the compute processor will have less impact on other applications, or on other simultaneous communication in the same application, and would perform better in systems where the network may be a bottleneck, e.g., workstation clusters.

5 Irregular subsets

While many applications may be able to specify an I/O transfer as a large, contiguous portion of a file, or perhaps as a regular pattern of smaller pieces of the file [NK95], there are some applications that need to request an irregular, discontiguous subset of a file. One example is the radio-astronomy data set described above. Another class of such applications includes computational quantum chemistry applications, which retrieve precomputed integrals from a table stored in a large file [Ken95]. Finally, many applications use sparse matrices, non-uniform mesh decompositions, or other irregular out-of-core data structures. How might disk-directed I/O support these kinds of applications?

5.1 Experiments

We simulate applications that request an irregular subset of blocks from the file, although we do assume that they could specify the entire subset up front in a list (e.g., in a batch request [NK95]). The list was in logical sorted order. In the disk-directed I/O system, each compute processor sent

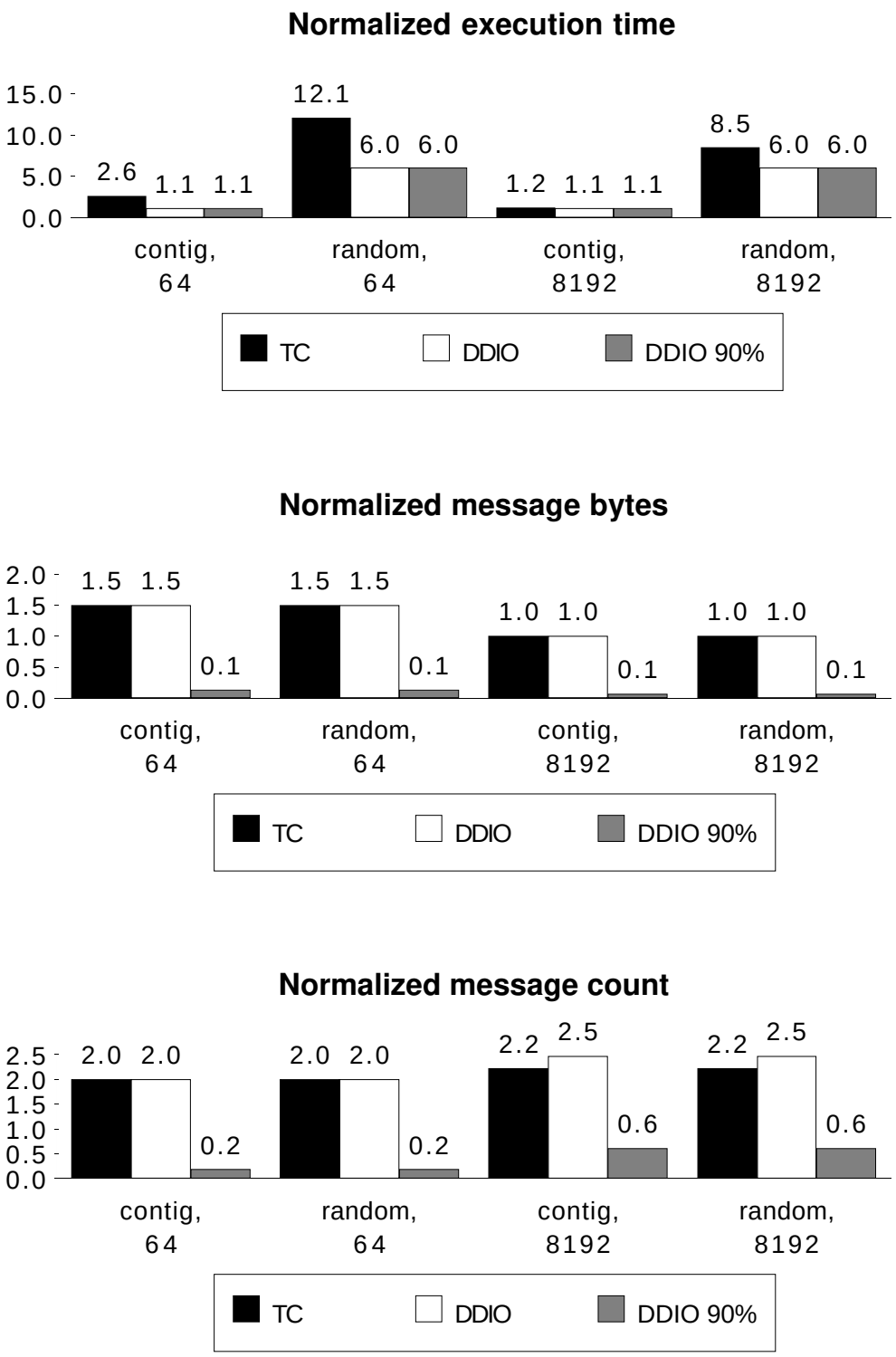


Figure 3: Results of data-dependent filtration experiments. Traditional caching is unchanged by filtering. We compare normal DDIO with a DDIO that filtered 90% of all records. The message metrics are normalized against the amount of traffic necessary with no filtration.

its list of requests to the appropriate I/O processors. The I/O processors converted the logical offsets to a list of physical block numbers, sorted the list for the best schedule, and processed the blocks much as before. In the traditional-caching system, each compute processor requested one block at a time, according to the list.

We used a 100 Mbyte file and selected an irregular subset of about 22% of the file blocks; of that subset, each processor requested those blocks that would also be distributed to it if the whole file were distributed cyclically, block by block, to all processors. Thus, this pattern was a subset of the `rc` pattern. The subset was arranged so that each processor received about the same number of blocks, and each disk received about the same number of read requests.

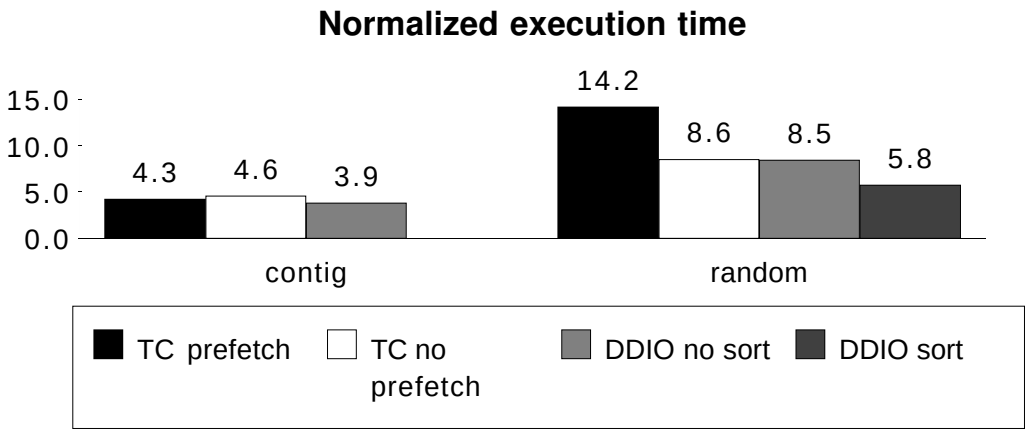
5.2 Results

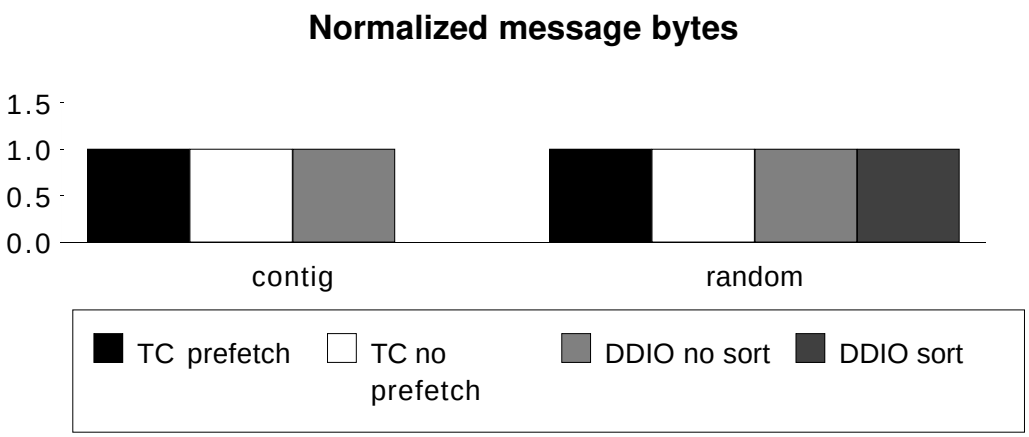
Figure 4 displays the results of these experiments. Traditional caching is presented both with and without the prefetching it normally uses. It is clear that in the random layout the prefetching made performance dramatically *worse*, by prefetching useless blocks on the mistaken assumption of sequential access, whereas in the contiguous layout those mistakes made little difference because those prefetches were quickly completed by the drive's own prefetching. A production system would need to include smarter prefetching policies [KE93, PG94].

Disk-directed I/O was faster than traditional caching, and obtained its usual benefit on a random layout due to sorting. It is no better or worse in terms of message traffic. The main point here is that disk-directed I/O need not be restricted to regular access patterns; as long as the request can be formed as a reasonably compact list of records to transfer, it works fine.

6 Conclusions

While our earlier work demonstrated the value of disk-directed I/O for large, collective requests, the examples were limited to regular structures with a data-independent distribution [Kot94]. In this paper we show that disk-directed I/O could accommodate irregularly structured requests, data-dependent distributions, and data-dependent filtering, with no loss in performance. Traditional caching, on the other hand, had costly prefetch mistakes on the irregular access pattern, doubled the network traffic when doing a data-dependent distribution function, and was not able to reduce the network traffic when doing data-dependent filtering. Although the difference in network traffic did not affect the execution time in our experiments, because we used a very fast network, it may affect performance on systems where the network is slower or shared (as in a workstation cluster).





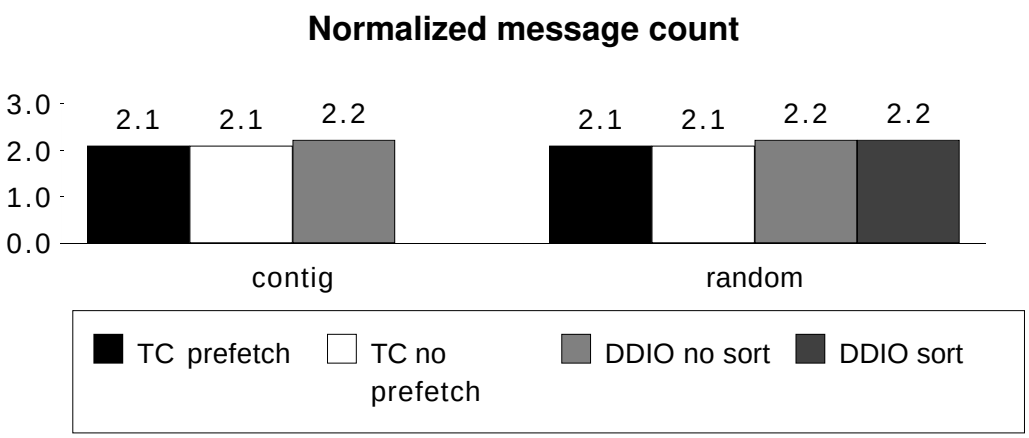


Figure 4: Normalized measures for experiments requesting an irregular subset of the file. Best execution time is computed from the peak disk bandwidth and the actual amount of data transferred. Best message count and message bytes are computed assuming that only the requested blocks will flow through the network.

Given that disk-directed I/O has the potential for all of these capabilities, the crucial remaining question is how to make these capabilities available to the programmer. In particular, how does the user (or compiler) tell the I/O processor about its distribution function, filtration function, and which file data to read? We are beginning to study this issue. We have proposed a new interface [NK95] that may play a role in specifying the desired subset of file data, and perhaps in data-independent distributions, but much more work remains to be done.

Availability

STARFISH is available at `http://www.cs.dartmouth.edu/research/starfish/` or by ftp to `ftp.cs.dartmouth.edu` in `pub/pario/STARFISH/`.

Many of these papers can be found via the URL `http://www.cs.dartmouth.edu/pario.html`.

References

[BBS+94] Robert Bennett, Kelvin Bryant, Alan Sussman, Raja Das, and Joel Saltz. Jovian: A framework for optimizing parallel I/O. In *Proceedings of the 1994 Scalable Parallel Libraries Conference*, pages 10–20. IEEE Computer Society Press, October 1994.

[BdC93] Rajesh Bordawekar, Juan Miguel del Rosario, and Alok Choudhary. Design and evaluation of primitives for parallel I/O. In *Proceedings of Supercomputing '93*, pages 452–461, 1993.

[BDCW91] Eric A. Brewer, Chrysanthos N. Dellarocas, Adrian Colbrook, and William E. Weihl. Proteus: A high-performance parallel-architecture simulator. Technical Report MIT/LCS/TR–516, MIT, September 1991.

[CF94] Peter F. Corbett and Dror G. Feitelson. Design and implementation of the Vesta parallel file system. In *Proceedings of the Scalable High-Performance Computing Conference*, pages 63–70, 1994.

[CFH+95] Peter Corbett, Dror Feitelson, Yarson Hsu, Jean-Pierre Prost, Marc Snir, Sam Fineberg, Bill Nitzberg, Bernard Traversat, and Parkson Wong. MPI-IO: a parallel file I/O interface for MPI. Technical Report NAS-95-002, NASA Ames Research Center, January 1995. Version 0.3.

[CFPB93] Peter F. Corbett, Dror G. Feitelson, Jean-Pierre Prost, and Sandra Johnson Baylor. Parallel access to files in the Vesta file system. In *Proceedings of Supercomputing '93*, pages 472–481, 1993.

[CK93] Thomas H. Cormen and David Kotz. Integrating theory and practice in parallel file systems. In *Proceedings of the 1993 DAGS/PC Symposium*, pages 64–74, Hanover, NH, June 1993. Dartmouth Institute for Advanced Graduate Studies.

[dBC93] Juan Miguel del Rosario, Rajesh Bordawekar, and Alok Choudhary. Improved parallel I/O via a two-phase run-time access strategy. In *IPPS '93 Workshop on Input/Output*

in Parallel Computer Systems, pages 56–70, 1993. Also published in Computer Architecture News 21(5), December 1993, pages 31–38.

[DdR92] Erik DeBenedictis and Juan Miguel del Rosario. nCUBE parallel I/O software. In *Eleventh Annual IEEE International Phoenix Conference on Computers and Communications (IPCCC)*, pages 0117–0124, April 1992.

[Dib90] Peter C. Dibble. *A Parallel Interleaved File System.* PhD thesis, University of Rochester, March 1990.

[DSE88] Peter Dibble, Michael Scott, and Carla Ellis. Bridge: A high-performance file system for parallel processors. In *Proceedings of the Eighth International Conference on Distributed Computer Systems*, pages 154–161, June 1988.

[EGKS90] Susanne Englert, Jim Gray, Terrye Kocher, and Praful Shah. A benchmark of NonStop SQL Release 2 demonstrating near-linear speedup and scaleup on large databases. In *Proceedings of the 1990 ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 245–246, May 1990.

[FBD94] Craig S. Freedman, Josef Burger, and David J. Dewitt. SPIFFI — a scalable parallel file system for the Intel Paragon. Submitted to IEEE TPDS, 1994.

[FPD93] James C. French, Terrence W. Pratt, and Mriganka Das. Performance measurement of the Concurrent File System of the Intel iPSC/2 hypercube. *Journal of Parallel and Distributed Computing*, 17(1–2):115–121, January and February 1993.

[GGL93] N. Galbreath, W. Gropp, and D. Levine. Applications-driven parallel I/O. In *Proceedings of Supercomputing '93*, pages 462–471, 1993.

[GL91] Andrew S. Grimshaw and Edmond C. Loyot, Jr. ELFS: object-oriented extensible file systems. Technical Report TR-91-14, Univ. of Virginia Computer Science Department, July 1991.

[HdC95] Michael Harry, Juan Miguel del Rosario, and Alok Choudhary. VIP-FS: A virtual, parallel file system for high performance parallel and distributed computing. In *Proceedings of the Ninth International Parallel Processing Symposium*, April 1995. To appear.

[HER+95] Jay Huber, Christopher L. Elford, Daniel A. Reed, Andrew A. Chien, and David S. Blumenthal. PPFS: A high performance portable parallel file system. Technical Report UIUCDCS-R-95-1903, University of Illinois at Urbana Champaign, January 1995.

[KE93] David Kotz and Carla Schlatter Ellis. Practical prefetching techniques for multiprocessor file systems. *Journal of Distributed and Parallel Databases*, 1(1):33–51, January 1993.

[Ken95] Rick Kendall. Personal communication. Pacific Northwest Laboratories, February 1995.

[KFG94] John F. Karpovich, James C. French, and Andrew S. Grimshaw. High performance access to radio astronomy data: A case study. In *Proceedings of the 7th International Working Conference on Scientific and Statistical Database Management*, September 1994. Also available as Univ. of Virginia TR CS-94-25.

[KHH+92] Masaru Kitsuregawa, Satoshi Hirano, Masanobu Harada, Minoru Nakamura, and Mikio Takagi. The Super Database Computer (SDC): System architecture, algorithm and preliminary evaluation. In *Proceedings of the Twenty-Fifth Annual Hawaii International Conference on System Sciences*, volume I, pages 308–319, 1992.

[KN94] David Kotz and Nils Nieuwejaar. Dynamic file-access characteristics of a production parallel scientific workload. In *Proceedings of Supercomputing '94*, pages 640–649, November 1994.

[Kot94] David Kotz. Disk-directed I/O for MIMD multiprocessors. In *Proceedings of the 1994 Symposium on Operating Systems Design and Implementation*, pages 61–74, November 1994. Updated as Dartmouth TR PCS-TR94-226 on November 8, 1994.

[KR94] Thomas T. Kwan and Daniel A. Reed. Performance of the CM-5 scalable file system. In *Proceedings of the 8th ACM International Conference on Supercomputing*, pages 156–165, July 1994.

[Kri94] Orran Krieger. *HFS: A flexible file system for shared-memory multiprocessors.* PhD thesis, University of Toronto, October 1994.

[LIN+93] Susan J. LoVerso, Marshall Isman, Andy Nanopoulos, William Nesheim, Ewan D. Milne, and Richard Wheeler. *sfs*: A parallel file system for the CM-5. In *Proceedings of the 1993 Summer USENIX Conference*, pages 291–305, 1993.

[Mas92] Parallel file I/O routines. MasPar Computer Corporation, 1992.

[MS94] Steven A. Moyer and V. S. Sunderam. PIOUS: a scalable parallel I/O system for distributed computing environments. In *Proceedings of the Scalable High-Performance Computing Conference*, pages 71–78, 1994.

[Nit92] Bill Nitzberg. Performance of the iPSC/860 Concurrent File System. Technical Report RND-92-020, NAS Systems Division, NASA Ames, December 1992.

[NK95] Nils Nieuwejaar and David Kotz. Low-level interfaces for high-level parallel I/O. Technical Report PCS-TR95-253, Dept. of Computer Science, Dartmouth College, March 1995. Submitted to IOPADS at IPPS '95.

[PEK+95] Apratim Purakayastha, Carla Schlatter Ellis, David Kotz, Nils Nieuwejaar, and Michael Best. Characterizing parallel file-access patterns on a large-scale multiprocessor. In *Proceedings of the Ninth International Parallel Processing Symposium*, April 1995. To appear.

[PG94] R. Hugo Patterson and Garth A. Gibson. Exposing I/O concurrency with informed prefetching. In *Proceedings of the Third International Conference on Parallel and Distributed Information Systems*, pages 7–16, September 1994.

[Pie89] Paul Pierce. A concurrent file system for a highly parallel mass storage system. In *Fourth Conference on Hypercube Concurrent Computers and Applications*, pages 155–160. Golden Gate Enterprises, Los Altos, CA, March 1989.

[Roy93] Paul J. Roy. Unix file access and caching in a multicomputer environment. In *Proceedings of the Usenix Mach III Symposium*, pages 21–37, 1993.

[SW94] K. E. Seamons and M. Winslett. An efficient abstract interface for multidimensional array I/O. In *Proceedings of Supercomputing '94*, pages 650–659, November 1994.

[TBC+94] Rajeev Thakur, Rajesh Bordawekar, Alok Choudhary, Ravi Ponnusamy, and Tarvinder Singh. PASSION runtime library for parallel I/O. In *Proceedings of the Scalable Parallel Libraries Conference*, pages 119–128, October 1994.